Project Number: 693349

State of the art approaches and tools

Giuseppe Airò Farulla, Pietro Braione, Marco Indaco, Mauro Pezzè, Paolo Prinetto

CINI, Aster SpA

Version 1.2 – 15/11/2017

Lead contractor: Universitat Pompeu Fabra
Contact person: Josep Quer Departament de Traducció i Ciències del Llenguatge Roc Boronat, 138 08018 Barcelona Spain Tel. +34-93-542-11-36 Fax. +34-93-542-16-17 E-mail: josep.quer@upf.edu
Work package: 3
Affected tasks: 3.2

Nature of deliverable[1]	R	P	D	O
Dissemination level[2]	PU	PP	RE	CO

[1] *R: Report, P: Prototype, D: Demonstrator, O: Other*

[2] ***PU**: public, **PP:** Restricted to other programme participants (including the commission services), **RE** Restricted to a group specified by the consortium (including the Commission services), **CO** Confidential, only for members of the consortium (Including the Commission services)*

COPYRIGHT

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 693349.

History of changes

VERSION	DATE	CHANGE	REVIEWER(S)
1.0	01/03/2017	Initial version.	Paolo Prinetto, Josep Quer, Jordina Sánchez Amat
1.1	25/07/2017	Added Sections 3.d and 6; Typos fixing; Adapted to the project's template.	Paolo Prinetto, Josep Quer, Jordina Sánchez Amat
1.2	15/11/2017	Document completely revised w.r.t. its original structure. In particular, hereby follows the most important changes made: • Overall confusion between software (that can be made or bought) and services has been removed; • Overall vagueness has been removed and now more clear indications about software and services to be employed in the project are given; discussions and agreements with the providers of these software and services have been made-are being made; • Removed overlap with deliverable D3.11; • The old general introduction about web application technologies has been removed; • The old general introduction about generic CMS systems for managing web pages has been reshaped and better specified to match the interest of the project; • The new Chapter 2 has been added to deal with CMS systems and state-of-the-art projects for managing multimedia content, introducing also the general principles inspiring our work; • The new Chapter 3 has been added to deal with state-of-the-art projects and web-based tools for developing front-end systems for managing forms and surveys; • The old Chapter 4 about the Analysis on multimedia content has been moved to be then merged to D3.11; • The new Chapter 4 about the GIS applications deals now with both software solutions and services; • The new Chapter 5 better summarizes the outcomes of the "build or buy" analysis identifying which component should necessarily be provided by the project to meet the requirements identified in D3.1and what instead can be provided by external services.	Pietro Braione, Mauro Pezzè, Jordina Sánchez Amat

INDEX

1. Introduction 5

2. Management and Distribution of Digital Content 6
   a. Introduction 6
   b. General Principles 7
   c. Content Management Systems 10
   d. Content Management Systems for Digital Content and Video 12
   e. Content Management Systems for Linguistic Disciplines and Preservation of Digital Assets 20
   f. European and International Platforms and Repositories 23
   g. Distribution of Digital Content 30
   h. Real-time Streaming of Video Content 31
   i. Considerations 35

3. Management and Distribution of Surveys 36

4. Standards for Archiving and Preservation of Digital Content 38
   a. Introduction 38
   b. Identification of resources 41
   c. Metadata 44
   d. Providers of Digital Preservation Services 48

5. Solutions and Services for GIS systems 49
   a. Introduction 49
   b. Geographic Information Systems 50
   c. GIS Client 53
   d. Web map server 54
   e. Web server 55
   f. Geodatabase 56

6. Design Considerations 57
   a. Management and Distribution of Digital Content 58
   b. Management and Distribution of Surveys 59
   c. Standards for Archiving and Preservation of Digital Content 60
   d. Solutions and Services for GIS Systems 62

7. Conclusion 63

1. Introduction

This deliverable report includes the results of the activities undertaken in WP3 *Digital Infrastructures*. The main objective of the WP3 is the development of the web platform that will host the contents to be produced by WP2 and WP4. Specifically, this document is connected to Task 3.2 (i.e., *State of the Art*).

The goal of the deliverable is to give an overview of existing software solutions that can be used for the platform development and whether they can be easily modified and integrated to fully respect the list of requirements defined in task 3.1 and the subsequent deliverable D3.1. As a result of this study a list of existing tools, software and web portals have been proposed and their performances with respect to the requirements defined in task 3.1. have been evaluated.

Basically, this deliverable presents the outcomes of a *build or buy* analysis that has been conducted in Task 3.2. By this analysis, the possibility of building the digital platform needed for the SIGN-HUB project on top of commercial, off-the-shelf (COTS) products or already existing international platforms has been in-depth evaluated. The core architectural requirements emerged from D3.1 have been carefully evaluated with respect to the existing technology and platforms. As a result, being the users' requirements quite demanding and the objectives of the project challenging, the need for custom software to be developed in Tasks 3.3 and 3.4 for building the SIGN-HUB digital platform has emerged; instead, the possibility to leverage on existing repositories and platforms to ensure long-term persistence and usability of digital content that will be provided either by the content providers (e.g., Work Package WP2) either by the end users through the SIGN-HUB digital platform is being evaluated. Clearly, the possibility to *buy* rather than to *build* existing technology, or the possibility of leveraging on existing services, would speed up the development activities in the SIGN-HUB project, but at the present state for some of its interfaces would also force to compromise and to fail in matching important requirements posed by D3.1.

The structure of the deliverable is the following:

- Section 2 deals with the management and distribution of digital content, both video and not, to be handled within the SIGN-HUB project and the requirements for data usability;
- Section 3 deals with the management and distribution of surveys, which are required in the SIGN-HUB project to let end users populate the Linguistic ATLAS with the descriptions of grammars other than the ones to be produced by the Work Package WP2;
- Section 4 deals with the standards for archiving and long-term preservation of digital content, both video and not, in terms of identification of resources, identification of information about the data, and well-known service providers in the field;
- Section 5 deals with software and services for handling geolocalization of digital content through Geographic Information Systems;
- Section 6 summarizes our design considerations;
- Section 7 finally briefly concludes this document.

2. Management and Distribution of Digital Content

a. Introduction

The aim of this section is to provide the reader with a throughout review of state-of-the-art solutions for the management and distribution of digital content, to evaluate the benefits of using cost effective and reliable content management system instead of using previous conventional website development methodologies and to identify if, and in case which, CMS should be used in which circumstances and in which conditions in order to subsequently select which best meet SIGN-HUB project needs. This section also summarizes the outcomes of our analysis and the actions we have already taken to meet the requirements of the SIGN-HUB project, as exposed in deliverable D3.1.

b. General Principles

Beyond proper collection, annotation, and archival, data stewardship includes the notion of *long-term care* of valuable digital assets, with the goal that they should be discoverable and re-usable, either alone, or in combination with newly generated data, for many years in the future, for research purposes that might even be still not foreseeable as of today[3]. The outcome from good data management and stewardship is therefore the collection of well-described high-quality that facilitate and simplify this ongoing process of discovery, evaluation, and reuse of the same data (clearly, restrictive licenses and requests for authorization may apply).

As a result of a joint collaborative effort from the scientific community to provide guidelines on enhancing the ability of both human individuals and machines to automatically find and use the data, the FAIR Principles have been designed[4]. They express that:

- data should be Findable;
- data should be Accessible;
- data should be Interoperable;
- data should be Re-usable.

By following the FAIR guidelines (i.e., a minimal set of community-agreed guiding principles and practices) and implementing a FAIR-compliant repository, researchers are confident that the outcomes of their activities are actually more easily discovered, accessed, interoperated, cited, and sensibly re-used.

At the core of the FAIR data formatting and publishing process, there is a comprehensive view on what constitutes data and how is it structured. In the following, we will define generically as *data* a comprehensive set of a data item with data elements, its metadata, and a unique and persistent identifier.

Having defined the basic concepts, the FAIR principles imply that:

- to be findable (F) or discoverable, data and metadata should be richly described to enable attribute-based search;
- to be broadly accessible (A), data and metadata should be retrievable in a variety of formats that are sensible to humans and machines using persistent identifiers;
- to be interoperable (I), the description of metadata elements should follow community guidelines that use an open, well defined vocabulary;
- to be reusable (R), the description of essential, recommended, and optional metadata elements should be machine processable and verifiable, use should be easy and data should be citable to sustain data sharing and recognize the value of data.

These FAIR Facets are obviously related, but technically somewhat independent from one another, and may be implemented in any combination, incrementally, as both data themselves as well as the knowledge of data providers evolve to increasing degrees of FAIR-ness. As such, the barrier-to-entry for FAIR data producers, publishers and stewards is maintained as low as possible, with providers being encouraged to gradually increase the number of FAIR Facets they comply with, as shown in Figure 1.

---

[3] *Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., Da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Nature, 3, 2016.*

[4] *Guiding principles for findable, accessible, interoperable and re-usable data, 2017, Force11, available on-line at https://www.force11.org/fairprinciples (accessed on 06/11/2017)*

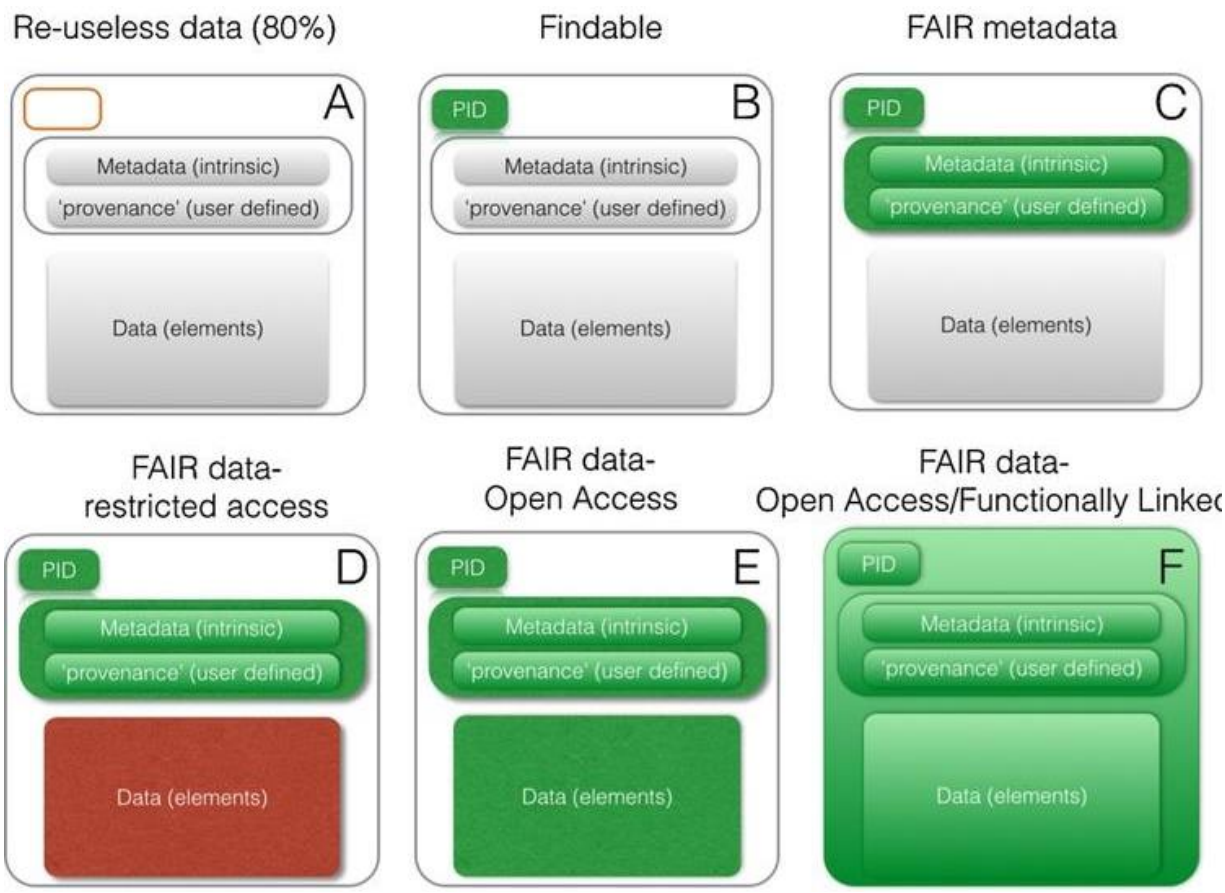

Figure 1 - Data structures as increasingly FAIR Digital Objects

With respect to the SIGN-HUB project, the data to be collected will be mostly (but not only) represented by video recording of interviews to signers, specifically targeting elderly signers and signers with language disorders. Other digital content will comprise, for instance, images of phoneme from Sign Languages, annotations, and documents about linguistic content. Such data will be of outmost importance not only to preserve the cultural heritage of a variety of Deaf communities spread all over Europe, but also to foster the development of new linguistic research activities in the (near) future. Therefore, it is highly recommendable for the SIGN-HUB project to comply with the FAIR principles since its beginning, by setting up a repository where data structures are stored while being easily findable, usable, retrievable and interconnected for research purposes. Thus, the project has to come up with solutions to manage data structures in their entirety: in this, it is of importance to deal with the definitions of metadata and global persistent unique identifiers and with the modern standards and tools to represent and use them. In fact, clearly putting data *on the web* is not enough. To be actually interoperable and reusable, data should not only be properly licensed, but also the methods to access and/or download them should be well described and preferably fully automated, using well established protocols.

The digital content to be produced within the SIGN-HUB project will concern also multimedia content posing ethical concerns restrictions that may conflict with its reusability, such as a large bunch of video interviews of signers with language disorders. Considering this requirement, the analysis given on existing platforms and repositories will deal also with the storage capacity they offer and with the capabilities they offer, or the methods that can be used, to ensure the protection of personal or other sensitive data. In any case, the management structure that is recommendable for the SIGN-HUB project to allow researchers to access to sensible information is:

- based on a careful Authentication and Authorization setup, with a clear procedure on who can access/request access to what (an audit trail is required to clarify such procedure);
- trust-based, so once a trustworthy researcher has signed in and accepted the terms of use, the digital platform should trust him/her thereinafter.

The metadata of a data item should be sufficiently rich that a machine or a human user, upon discovery, can make an informed choice about whether or not it is appropriate to use that data item in the context of their analysis; this metadata should be machine-readable to facilitate automated data harvesting while maintaining proper attribution. Metadata should inform the consumer about the license and the restrictions (if any) that apply to the data item: it should inform about any access-control policy, such that consumers can determine which components of the data they are allowed to access. The Metadata within the Data Object should inform about the authentication protocol leading to access, if applicable.

Concerning Identifiers, there are ongoing and fierce debates on what exactly constitutes a persistent identifier (often referred to as PID). We propose to allow many identifiers in a FAIR data publishing environment as long as any identifier is uniquely referring to only one concept and the publisher can provide a clear policy and description on the maximum achievable guarantee for persistent resolving of the identifier to the correct location/meaning. Obviously, locally used identifiers that cannot be mapped automatically to community adopted and publicly shared identifier schemes are not FAIR. The data publisher choosing a proprietary identifier scheme, will need to provide appropriate and correct mappings to public identifiers to be considered FAIR.

Metadata (as detailed later in Section 4.c Metadata) for the digital content to be produced within the SIGN-HUB project should therefore include at least the following information:

- persistent identifier;
- author;
- title;
- date;
- data type;
- data origin;
- data format;
- data weight.

c. Content Management Systems

A Content Management System (CMS) is a particular kind of web application that is oriented toward the creation, management and publishing of content (e.g. images, videos, documents).

The Content Management Systems make web page content management simple for even non-technical users, allowing designated users to change most aspects of their sites themselves, without the help of a web developer. These systems typically endow content administrators and consumers (i.e., regular users that just browse the website) with a set of relevant aspects such as modularity, independence between content and its presentation, access management, user control, or configurable visual appearance and layout of content[5].

Figure 2 - CMS Architecture

CMSs store the data in databases to make maintaining a website less cumbersome as there are no separate files for each website page. As the content is pulled dynamically from the database it is very easy to control the placement of it on the website by setting visibility rules in advance.

A CMS separates the content from the User Interface (UI) designed by using a theme layer which is responsible for rendering page layout and CSS styles associated with it. To achieve that separation from the user interaction perspective, the Content Management Systems often provide standardized *Back-end* and *Front-end* interfaces. The former is the administrative view of the CMS, while the latter is the end-user facing view.

A CMS is now the preferred choice for those who want easy content changes, simplified control of large amounts of content, a choice of plug-ins to accomplish a wide variety of tasks. In the following a comparative analysis of the most widely used CMSs for managing digital content, especially (but not only) video, is presented.

[5] *Joao Paulo Pedro Mendes de Sousa Saraiva, Development of CMS-based Web Applicationswith a Multi-Language Model-Driven Approach 2013 (Ph.D. Thesis)*

Our analysis covers the following aspects:

- *Content Management:* this feature analyzes which management approaches is used by the CMS systems. CMS typically use one of two management approaches: page-centric and content-centric. A page-centric approach considers that the website's structure (i.e., a set of pages and components) must be defined first, and afterward content is defined within the context of that structure. On the other hand, a content-centric approach dictates that the content itself must be defined, and afterward the administrator can specify a structure of pages that will show some of that content.

- *Content Creation:* this feature determines in which way the CMS system allows its administrator users to perform the acquisition (gather the content from some existing source), aggregation (edit the content, divide it into components, and augment it to fit within the desired metadata system) and authoring (create the content from scratch) of contents.

- *Content Publishing*: this feature determines whether the CMS system allows its administrator users to customize the website's structure, in such a way that effectively allows visitors to perceive the website's organization as a structured hierarchical set of pages. Besides the possibility of customizing the website's structure, moreover this aspect determines whether administrators should also be able to customize the website's visual layout (i.e., the website's look-and-feel, such as the colors used, or the relative location of each container that pages will use to show content).

d. Content Management Systems for Digital Content and Video

Of particular interest for the SIGN-HUB project, there is the need for the project's digital platform to host, manage, and retrieve multimedia content, and particularly videos. In fact, the platform will host two front-end interfaces dedicated to facilitate data entry (including multimedia file upload from content creators) and to allow streaming of the multimedia content uploaded (with particular regards to videos and documentaries to be created in task T2.4).

In addition, the platform should provide content creators with the possibility to upload videos to the repository in a fluid process and users and researchers with the possibility to access them easily on multiple platforms and devices. Indeed, this chapter deals with video content management systems, or video CMS: software that enable to centralize, manage, and deliver video online.

Videos generated in the project will create the corpora of Sign Languages grammars but also will relate to private lives of people coping with disabilities (e.g., not hearing) and/or language disorders. For this reason, agreeing with guidelines provided by Deliverable D1.5 "Data Management Plan", this analysis excludes a priori the possibility for the SIGN-HUB project to leverage on commercial public video platforms such as YouTube (www.youtube.com) or Vimeo (www.vimeo.com). In fact, while the available privacy settings on these video sites help minimizing the risk of inadvertently sharing personal information, still this risk is far too great and not compliant with the dedicated European laws and regulations on the purpose. Also, these commercial platforms do not seem to adhere to the FAIR principles.

Video CMSs store the data in databases to make the maintenance of a website less cumbersome, as there are no separate files for each website page. As the content is pulled dynamically from the database it is very easy to control the placement of it on the website by setting visibility rules in advance.

The usage of a video CMS should come with benefits for the SIGN-HUB project, being these benefits for instance the support for large high-definition, even uncompressed, video files, the optimized video delivery for streaming on every viewer's device (e.g., on-line conversion to the most proper format and video coded), the possibility to adapt bitrate streaming on the actual speed of the Internet connection of the viewer, the support for mobile, and an innate support for security and data privacy.

Central to the analysis given in this chapter on the possibility to leverage on off-the-shelf state of the art video CMS for the purposes of the SIGN-HUB project have been the reports on "How to Take Video Mobile with Enterprise Video Content Management"[6] and on "Reviews for Enterprise Video Content Management"[7] from Gartner, widely recognized as one of the most influential market analyses. This last report defines video CMS as "software, appliances, or software as a service (SaaS) intended to manage and facilitate the delivery of one-to-any, on-demand video across Internet protocols". In particular, the recent "magic quadrant" analysis from Gartner, shown in Figure 3, has represented a valid starting point for the analysis given in this section. Findings from Gartner have also been compared with the ones from a similar report by Forrester[8], which identifies the most significant providers in the category and researched, analyzed, and scored them.

---

[6] *Whit Andrews, How to Take Video Mobile with Enterprise Video Content Management, 2013, Gartner, available on-line at https://www.gartner.com/doc/2516615/video-mobile-enterprise-video-content (accessed on 03/11/2017)*

[7] *Whit Andrews and Adam Preset, Reviews and Magic Quadrant for Enterprise Video Content Management, 2016, Gartner, available on-line at https://www.gartner.com/reviews/market/enterprise-video-content-management (accessed on 03/11/2017)*

[8] *Philipp Karcher, Stephen Powers, Steven Kesler, and Danielle Jessee, Enterprise Video Platforms and Webcasting, 2015, Forrester, available on-line at https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Video+Platforms+And+Webcasting+Q1+2015/-/E-RES117998 (accessed on 03/11/2017)*

Figure 3 - Magic Quadrant for Enterprise Content Management Systems for Video

The Magic Quadrant above-mentioned includes technologies developed by 15 vendors that provide, in the European or American market, solutions for managing the workflow, storage, search and integration of video content. These vendors are classified as Leaders (who have developed flexible, extensible products that are effective in a variety of use cases), Challengers (who do not fully qualify as Leaders although they have a defensible business position and have demonstrated a commitment to the market), and Niche Players (who are developing the necessary capability to pursue enterprise video content management opportunities more fully after having targeted other markets).

In the following, a list of technologies for video CMS is given. However, it is also important to state that the digital content to be produced in the SIGN-HUB project is not limited to video: documents and images about linguistic structures (e.g., phoneme) and heritage will be collected too. Thus, technologies and repositories in the following analyses are to be assessed also considering whether they can cope easily with digital content that is not video.

Considering enterprise solutions, Brightcove (www.brightcove.com) is the top market leader solution for video CMS. It provides users with functionalities to save, sort, and search the video library, link the videos to custom metadata, manage multiple accounts and their privileges (e.g., by giving or revoking the possibility to upload videos to content providers), go for live streaming, manage playlists, and support multiple devices for streaming, here included on-line

transcoding for better visualization on tablets and mobile devices. In addition, it is compatible with popular CMS services such as WordPress, Drupal, and Joomla.

Panopto (www.panopto.com) is a video CMS specifically designed for video management, live streaming, analytics, and mobile. Basically, implements what is called an *enterprise version of YouTube*, giving access to a centralized video library, that can be stored in the premises or in the cloud, for video upload, transcoding (Panopto automatically formats every video for optimal playback), streaming (even on mobile screens), and analytics. Search among videos is implemented not only on their metadata, but also through video content analysis techniques (search inside videos).

Kollective (kollective.com) has developed Software-Defined Enterprise Content Delivery Network (SD ECDN), a network for worldwide digital distribution of media content that is based on Panopto to manage all of its functionalities needed for live streaming, recording, managing, and sharing video across its customers. The network is based on the cloud and thus attractive only for projects that do not need to store video content on premises.

Similarly to Panopto, ViMP (www.vimp.com), acts like a YouTube platform for corporations functionally, still offering additional functionalities, such as the possibility to store the database on premises, the automatic conversions of videos uploaded into web-capable formats, including those for mobile devices, advanced searching functionalities, support for groups and large number of users.

MediaPlatform (www.mediaplatform.com) is the producer of PrimeTime, a corporate video CMS with dedicated social computing capabilities and support for automatic transcoding, mobile streaming, managing security and authorization profiles, and for maximising through advertising the reach of its video communications. PrimeTime also offers the ability to perform a system-wide or asset-specific search across all media content leveraging on innate capabilities for content analysis.

Kaltura (corp.kultura.com) is a video CMS which focuses on providing tools to assist content providers in creating value and effective digital communications with videos. It has all the functionalities needed to implement streaming also live, online transcoding, support for tablets and mobile devices, and support for video analytics and management of security features.

VBrick (vbrick.com) has developed Rev, a cloud product (now bought by CISCO) comprising a video CMS specialized in delivering video solutions tailored for large enterprises. It implements a cross-platform high-quality, reliable, and secure video solution. Its video streaming capabilities leverage on universal IP streams and CISCO technology.

Polycom (www.polycom.com) identifies enterprise video content management and collaboration as a key strategic target market. About one in four of the customer references identified by Gartner for their Magic Quadrant analysis (other than those provided by Polycom itself) had considered Polycom for their projects.

Qumu (www.qumu.com) has developed Qx, an enterprise video platform built to solve real-world business video challenges: it offers solutions related to Unified Communications, Social Business, Speed to Market, Executive Address and Training Delivery, and it seamlessly integrates with existing platforms like MS Sharepoint, Yammer, Jive, and IBM Connections.

Haivision (www.haivision.com) is a cloud-based solution to provide workflows for publishing and distribuiting videos, with support for Video content management, easy publishing on social media platforms, online transcoding (videos are optimized to use as little bandwidth as possible), analytics & reporting tools to track performance.

Vidizmo (www.vidizmo.com) is a corporate video CMS designed to empower organizations for using video and media content in any format for corporate communication, marketing and training. It supports the delivery of webcasts, webinars, and live event broadcasts via live video streaming. It includes facilities for seamless device detection and a mobile user interface for mobile access and video streaming.

Agile Content (www.agilecontent.com) has developed a platform specifically designed for business, particularly for maximising return on investments by companies based on Big Data

analysis through Machine Learning techniques for audience profiling and advanced segmentation. This platform includes tools for Video Management and Distribution that unify the management of video, audiences, and monetization (advertising) while providing support for on-demand and live consumption of media content via multiple devices. Finally, the Agile platform has a strong facility for the delivery of video content across the internet.

IBM (www.ibm.com/cloud-computing/solutions/video/) has developed a Cloud Video platform that targets large enterprises by providing them with capabilities for on-demand video management, monetization opportunities, and maximization of viewer engagement. The technology behind the platform has been developed by Ustream, an IBM company and part of the IBM Cloud Video division.

The solution provided by Genus Technologies (www.genusllc.com), named Unified Media Management, aims at unifying the management of enterprise content to simplify the process of creating, capturing, managing and sharing digital assets and content. Concerning to video content, it is based on IBM technology and integrated with IBM software.

KZO Innovations (kzoinnovations.com) has developed a Video Suite for an enterprise-focused Content Management. It is particularly focused on using on-demand video to allow employees to share information and knowledge in a collaborative and asynchronous environment. KZO offers integration to a very broad set of collaboration applications for playback and administration, and its player is designed to serve as a means of allowing workers to collaborate on video objects and use videos in ways others vendors do not, such as rich annotation.

Sonic Foundry (www.sonicfoundry.com) has developed a Mediasite Video Platform as an automated and scalable system for creating, publishing, searching and managing video content. It offers advanced capabilities for both video streaming and publishing (support for mobile and live streaming, video search, engagement tools and analytics) and for video management (innate support for collaborative editing and security functionalities). It was an early leader in video search and has a strong and well-established business in lecture capture.

Other video CMS are more focused on their capabilities to target the right users for enterprises rather than on these needed to store and manage the video content per se. For instance, Ooyala (www.ooyala.com) is a video CMS specifically designed for digital TV providers. It includes functionalities needed to maximize the return for any video business: the support for managing the complete video content management workflow, publishing to major broadcasters, as well as analytics capabilities for advanced business intelligence.

All of these CMSs expose functionalities of large interest for the purposes of the SIGN-HUB project, including, but not being limited to:

- Video Uploading/Download, also in many format and supporting also high-quality lossless codecs;
- Large video file uploads via FTP;
- Expanded video metadata set;
- Video Transcoding and Playback, even leveraging on Embedded HTML5, both in high and low resolution;
- Automatic thumbnail extraction;
- Adaptive UI for mobile;
- Integration with Universal Subtitles - viewing and adding annotations via video player;
- Content licensing including Creative Commons licensing, the GNU Free Documentation License (GFDL).

Nevertheless, these solutions seem to be also considerably expensive and, being proprietary and targeting principally the needs of (very) large enterprises, it is also not clear how they

threat intellectual propriety of the digital content and which services (e.g., APIs) they may expose to allow an easy integration with any digital platform (e.g., to enable video streaming).

Considering instead open source alternative solutions for video CMSs, Opencast (www.opencast.org) is a free and open source solution for automated video capture and distribution at scale, specifically tailored for the need of universities and research institutions worldwide. In particular, it provides a scalable infrastructure for encoding and enriching video with metadata, preview images, brands, captioning and text analysis to make the media more discoverable and accessible, with powerful search capabilities for static, dynamic, and user-generated metadata. The Opencast player can be used as a standalone application, or embedded inside other applications like blogs, wikis or content management systems. Opencast Playback enables slide segmentation and in-video text search. All player functionality is fully accessible, supporting assistive technology across multiple platforms.

MediaDrop (mediadrop.video) is a media-oriented content manager previously known as "MediaCore CE". It aims to make the task of publishing audio and video digital content simple and easy. It is a modular video, audio, and podcast publication platform which can be extended with plugins. It is designed to give a high degree of control over the resources presentation and administration, as well as tracking statistics and adding comments. MediaDrop is a cross-platform CMS. It is easy to browse video or audio files, add and upload them in one click, manage media library via a sophisticated administrative interface. Video encoding is handled by a corresponding automated plugin. And moderation is pleasant using a nice admin review system. Published media has comments, views, and likes that help CMS in ranking the most popular items. MediaDrop also has social media sharing support, so users can comment, embed, or share video and audio through social network.

CumulusClip (cumulusclips.org) is a free and easy to use video sharing script to build video websites and repositories. CumulusClips video sharing script produces video compatible on iOS & Android mobile devices, as well as all the major browsers for desktop PCs.

Plumi (plumi.org) is a video-sharing web application, totally free and open-source, produced by EngageMedia. It groups features and tools to create video-sharing applications and repositories, such as server-side transcoding of most video formats, upload progress bar, thumbnail extraction, HTML5 video playback and embedding, support for subtitles, large file uploading via FTP, social media integration, threaded commenting and user feedback forms, customised user profiles and a range of other useful features. The software includes a wide array of functionality to facilitate video distribution and community creation.

ClipBucket (clipbucket.com) is an open-source and freely downloadable PHP script to help researchers quickly and easily build their platform for video-sharing, also comprising interaction with social networks, such as direct and private messaging, creation of groups and playlists, and photo and images sharing. ClipBucket is not limited to just videos, but it's a complete Multimedia Management tool to manage videos, photos and audios all together on one single platform. ClipBucket supports a responsive design, so that the web pages created and maintained through it look good on all devices (desktops, tablets, and phones), an image management application focused on organizing digital images, and on-the-fly FFMPEG-based video conversion supporting direct streaming through HTML5 Players even starting from uncompressed files or anyway video not-directly stored in a format suitable for video streaming.

Video Share Video on Demand (VOD, videosharevod.com) is a plugin for Wordpress, arguably the most spread and used Content Management System for web pages worldwide. It is basically a solution for sharing on demand video. It includes a plethora of features like AJAX multi uploader, display and update of video list, show video preview in list, mobile video upload and playback, HTML5/RTMP/HLS player. It supports live video broadcasting, live video streams on site pages, and setup membership for accessing or broadcasting live video.

All of these open source CMSs expose functionalities of large interest for the purposes of the SIGN-HUB project, including, but not being limited to:

- video uploading and download;

- large video file uploads via FTP, even considering massive uploads;
- expanded video metadata set;
- video transcoding and playback, even leveraging on HTML5, both in high and low resolution;
- automatic extraction of thumbnail and information (e.g., duration, resolution, bitrate, and file size);
- adaptive UI for mobile;
- creation of playlists;
- definition of role playlists: assign videos as accessible by certain roles;
- integration with universal subtitles - viewing and adding annotations via video player.

Nevertheless, problems may arise since these open source CMS typically lack of codecs that are instead required to convert videos for streaming purposes; in addition, the conversions require important resources like CPU time, memory, long process time, typically not available on budget shared hosting. For the purposes of the SIGN-HUB project it is thus recommendable to refer directly to large repositories and platforms, possibly deriving from other research projects in the area of digitalizing humanities content, that embed capabilities for video management and distribution.

Of relevance for this document and overall for the SIGN-HUB project are the tools and software, specifically open source ones and mainly web-based, that may represent valid alternatives for the Digital Management Assessment (DAM): management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval, and distribution of digital content. DAM also refers to the protocols for downloading, renaming, backing up, rating, grouping, archiving, optimizing, maintaining, thinning, and exporting video or other digital content. Digital management is based on content models that represent data objects (units of content) or collections of data objects. The objects contain linkages between data streams (internally managed or external content files), metadata (inline or external), system metadata (including a PID – persistent identifier – that is unique to the repository), and behaviors that are themselves code objects that provide bindings or links to disseminators (software processes that can be used with the data streams). Content models can be thought of as containers that give a useful shape to information poured into them: if the information fits the container, it can immediately be used in predefined ways.

Generally, the asset being managed is collected and stored in a digital format. There is usually a target version – referred to as "essence" – generally the highest-resolution and highest-fidelity representation. The asset is detailed by its metadata. Metadata is the description of the asset and the description depth can vary depending on the needs of the system, designer, or user. Metadata can describe, but is not limited to, the description of:

- asset content (what is in the package?);
- the means of encoding/decoding;
- provenance (history to point of capture);
- ownership;
- rights of access;
- as well as many others.

There exist some predefined standards and template for metadata, the most relevant for the project being detailed in the following. In cases of systems that contain large-size asset essences, there are usually related *proxy* copies of the essence. A proxy copy is a lower-resolution representation of the essence that can be used as a reference in order to reduce the overall bandwidth requirements of the DAM system infrastructure. It can be generated and retained at the time of ingestion of the asset simultaneous or subsequent to the essence, or it can be generated on the fly using transcoders. Both these options are to be evaluated while

choosing services to possibly include in the SIGN-HUB project, for which the real-time streaming of large video content is a strong need.

Concerning DAM systems specifically tailored for the preservation of digital content, Concerto (concerto.sourceforge.net) is a preservation and collection oriented open source DAM, but it does have some features which potentially make it suitable for general use too. It is based on PHP and MySQL and leverages on the GPLv2, which presents few issues for third party developers (although integration with other solutions is subject to the usual restrictions with any GPL code). Concerto uses an editor/viewer configuration. Viewers can be delivered via a low-profile JavaScript application which accesses the Concerto data. Concerto has other less specialized features including: zoom image previews, custom metadata schemas, automated proxy generation (thumbnails and previews), batch importing and embedded metadata reading, LDAP integration and what they call *hierarchical authorizations*, which sounds like workflow.

EnterMedia (entermediasoftware.com) represents a good solution for audio and video digital content. It is Java based, so mainly suited to the needs of enterprises, and leverages on GPL license (however it is not clear to which version it refers). Formerly OpenEdit DAM, EnterMedia is an open source Digital Asset Management system developed using the OpenEdit content management framework. It includes full support for the typical range of facilities that modern DAM systems should include as standard, including extraction of embedded metadata, bulk uploading, transformation of image based assets etc. By default, EnterMedia uses XML files rather than a database, however, database connectors are available for those who are not keen on this approach. The OpenEdit framework is well established and EnterMedia's use of it is as well as Java marks it as suitable for enterprise use.

Considering repositories to store and preserve video content that can be later accessed and re-used, the following three frameworks expose features that are of interest for the SIGN-HUB project, and that might even be leveraged to build a SIGN-HUB custom repository.

Dspace (www.dspace.org) is a framework for developing Digital Asset Management solutions. Developed by HP and MIT Libraries, it is used extensively by academic and research organizations which makes it well suited for preservation usage scenarios. DSpace implementations are organized into communities which have responsibility for collections which are in turn composed of assets. DSpace is highly configurable and includes a flexible workflow for applying metadata to assets that will suit complex metadata. It is based on Java and leverages on the BSD license.

Fedora Commons (Flexible Extensible Digital Object Repository Architecture, www.fedora-commons.org), a repository for digital content that has focus on preservation, is as well based on Java and leverages on the BSD license. It may represent a valid tool for the underlying architecture for a digital repository, and is not a complete management, indexing, discovery, and delivery application. It is a modular architecture built on the principle that interoperability and extensibility are best achieved by the integration of data, interfaces, and mechanisms as clearly defined modules. Fedora supports two types of access services: a management client for ingest, maintenance, and export of objects; or via API hooks for customized web-based access services built on either HTTP or SOAP. A Fedora Repository provides a general-purpose management layer for digital objects, and containers that aggregate mime-typed datastreams (e.g., digital images, XML files, metadata). Out-of-the-box Fedora includes the necessary software tools to ingest, manage, and provide basic delivery of objects with few or no custom disseminators, or can be used as a backend to a more monolithic user interface. Fedora supports ingest and export of digital objects in a variety of XML formats. This enables interchange of objects between Fedora and other applications, as well as facilitating digital preservation and archiving.

Concerning entirely web-based pure DAM systems, Notre DAM (notredam.org) seems so represent a less complex alternative to DSpace or Fedora, also well worth checking out for non-academic use. It is based on Python and MySQL, and leverages on the GPLv3 license. It has been developed by CRS4 (Center for Advanced Studies, Research and Development in Sardinia) and although it has an academic background, it is somewhat simpler than Fedora or

DSpace and shares a number of characteristics with more commercially oriented open source DAM systems. Notre DAM uses the MediaDART framework which provides a number of media processing features that are ideal for Digital Asset Management. Notre DAM is also tightly integrated with XMP and contains a number of options for XMP based metadata manipulation. The application itself contains the core fundamentals required for serious DAM including a web based interface, support for images, video and documents, workflows, multiple taxonomies and a variety of other functionality such as Geotagging.

Islandora (islandora.ca) is an open source digital repository system based on Fedora Commons, including however also the well-known CMS for web pages Drupal and a host of additional applications. It is open source software (released under the GNU General Public License) and was developed at the University of Prince Edward Island by the Robertson Library. Islandora releases *solution packs* which empower users to work with data types (such as image, video, and pdf) and knowledge domains (such as digital humanities). It may be used to create large, searchable collections of digital assets of any type and is domain-agnostic in terms of the type of content it can steward. It has a highly modular architecture with a number of key features, including for instance:

- support for any file type (via the Fedora repository system);
- multi-language and functionality support via Drupal;
- a modular Solution Pack framework for defining specific data models and associated behaviors, including standard Solution Packs for audio, PDF, images, paged content, videos, and web archives;
- support for any XML metadata standard, including unique schemas;
- a formbuilder module which allows the creation of a data-entry/editing form for any XML schema;
- editorial workflows for approving submissions to the repository.

Islandora seems the most promising solution among the ones detailed in this analysis: it is thus recommended for the SIGN-HUB project to leverage on it to build a dedicated repository for its digital content and video if no existing platform can be found that already matches the requirements of the project (e.g., in terms of allowing real-time streaming and requiring authorization for display some of the content).

e. Content Management Systems for Linguistic Disciplines and Preservation of Digital Assets

The SIGN-HUB digital platform will expose a first interface for Sign Language grammars and a second interface for the Linguistic ATLAS. These will comprise, among other features, a hybrid system of text, images and video digital content to support navigation of deaf signers, exposing features such as interactive maps, unconstrained search feature combination search, an icon-based system to search for phonemes (e.g. handshapes), and searchable examples.

In this, capabilities for creating, storing and making accessible linguistic structures and grammars will be of paramount importance. This section deals with existing technologies and platforms that represent the state of the art for web-based tools to create, delivery, and manage Linguistic ATLAS.

In the following, a list of technologies and state of the art European projects that are focused on linguistic data is given.

WALS (wals.info), or the World Atlas of Language Structures (WALS) is a database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials. It is released on the Internet and maintained by the Max Planck Institute for Evolutionary Anthropology and by the Max Planck Digital Library. Properties are grouped in Chapters and represented as features. Combining up to four features contemporaneously, the user can crawl the database, comprising of structures from 2679 different languages spoken worldwide, to search for corresponding languages. The ATLAS provides information on the location, linguistic affiliation and basic typological features of a great number of the world's languages. The information of the ATLAS is published under a Creative Commons license.

DOBES (dobes.mpi.nl/dobesprogramme/), or DoBeS, is an acronym for an international organization and project with the German name Dokumentation bedrohter Sprachen ("Documentation of Endangered Languages"). This organization seeks to archive languages due to an expectation that many if not most of the world's languages (currently about 7,000 languages) will diminish and disappear by the end of the century. Language Documentation is the reaction of the linguistic community to the immanent disappearance of the majority of the world's languages. Its main goals and the unique nature of each documentation team and setting shape the documentary record that is submitted to the digital archive. Each documentation is carried out in close cooperation with the speech community. Audio and video data from a variety of genres are collected. The data are described with a set of standardized metadata categories and digitally archived according to open standards and made accessible. In addition, the archive has to take care of the long-term persistency of the digital material. The DOBES archive is organized into smaller subarchives, one per DOBES project. From its beginning, the DOBES programme wanted to take advantage of modern state-of-the-art technology, and where necessary drive technology to suit the needs of the documentation work. Therefore, the following topics were discussed and widely agreed upon, in particular in the pilot phase:

- specifications for archival document formats to promote long-term accessibility;
- recommendations for recording and analysis formats, and tools to ensure quality and reduce the conversion effort;
- the creation of new tools that support the audio/video annotation work, the metadata creation and the navigation in metadata domains, advanced web-based frameworks to access and enrich archived resources.

Many other details on the DOBES project, the quality standards it has set, and its achievement can be found for instance in the reference paper: http://www.mpi.nl/lrec/2002/papers/lrec-pap-02b-dobes-talk-final.pdf .

Terraling (www.terraling.com) is a collection of searchable linguistic databases that allows users to discover which properties (morphological, syntactic, and semantic) characterize a language, as well as how these properties relate across languages. This system is designed to

be free to the public and open-ended. Anyone can use the database to perform queries. Linguistic researchers can put up their own data using the open-ended framework provided.

In all of these databases, queries on the languages and grammars are highly constrained (e.g., WALS includes a search tool that allows to list all the languages in the database which share only up to four values for specific grammatical features, while SIGN-HUB aims at providing unconstrained exploration of languages and grammatical features) and provided only through a limited user interface, hard to follow for non-expert users. Therefore, although some of their features are inspiring and could be implemented in the front-end solutions for the project, we already see that they could not be used for the project as a back-end solution.

In addition, we have analyzed how the following projects and corpora of Sign Languages that have addressed issues similar to the ones posed by the SIGN-HUB project, to evaluate whether similar solutions may be applied:

The NGT corpus (Radboud University Nijmegen, www.ru.nl/corpusngten) grouped and recorded 92 Deaf signers from all over the Netherlands. Some clips have been added with voice over (leveraging on interpreters) and subtitles (leveraging on ELAN annotations). The corpus is released free of use both for video streaming and downloading. The license enforced is the CC and no detailed information is given about the open access policy. The data (clips and other information) have been stored through the Max Planck Institute (MPI) for Psycholinguistics in Nijmegen, which has a great expertise in the field of large collections of linguistic data. The language archive of the MPI institute (tla.mpi.nl/resources/archiving-service) seems adequate for the purposes of the SIGN-HUB project; it has been contacted to verify it meets the security requirements posed by the project itself (e.g., restrictions should apply on some of its data).

The BSL corpus (Economic and Social Research Council, http://www.bslcorpusproject.org) grouped and recorded UK Deaf signers. Some clips have been added with voice over (leveraging on interpreters) and subtitles (leveraging on ELAN "annotations"). The corpus is released free of use both for video streaming and downloading. The license enforced is the the Creative Commons Attribution-ShareAlike 4.0 International License and the open access policy is detailed on-line: http://www.bslcorpusproject.org/cava/open-access-data/. In particular, viewing or downloading the Open Access (narrative or lexical elicitation) video data first requires agreeing to some terms and conditions. Viewing or downloading the Restricted Access (interview or conversation) video data, requires a user license. The data (clips and other information) have been stored through the CAVA (human Communication: An Audio-Visual Archive) Repository, which is a digital video repository to support the work of the international human communication research community. It contains rights-cleared primary audio and video recordings made by UCL's human communication researchers and their associates. The CAVA repository (http://www.ucl.ac.uk/ls/cava/faq.shtml) seems adequate for the purposes of the SIGN-HUB project; it has been contacted to verify it meets the security requirements posed by the project itself (e.g., restrictions should apply on some of its data) and to apply for a license.

The LSFB corpus (University of Namur, http://www.corpus-lsfb.be/index.php?lang=En) grouped and recorded Francophone Belgian Deaf signers for a total amount of 150h of recording. Some clips have been translated to French. The corpus is released free of use for video streaming. The license enforced is the Creative Commons Attribution-ShareAlike 4.0 International License and no detailed information is given about the open access policy. The data (clips and other information) have been stored through a custom repository, developed with the same technologies envisaged to develop the SIGN-HUB platform. There is no indication that this repository is adequate for the purposes of the SIGN-HUB project.

The AUSLAN corpus (The Hans Rausing Endangered Languages Project, http://www.auslan.org.au/about/corpus/) grouped and recorded 100 native and near-native Deaf signers from Australia (for total 300 hours of digital videotape). Some clips have been added with subtitles (leveraging on ELAN "annotations", http://new.auslan.org.au/about/annotations/). The corpus is released free of use for video streaming. The license enforced is the CC and detailed information is given about the open access policy at the following link: https://www.soas.ac.uk/elar/using-elar/access-protocol/ .

The data (clips and other information) have been stored through the Endangered Languages Archive (ELAR) at SOAS University of London (https://elar.soas.ac.uk/Collection/MPI55247). The ELAR language archive seems adequate for the purposes of the SIGN-HUB project but should be contacted to verify if it meets the security requirements posed by the APDCAT authority and to fulfil a license (https://www.soas.ac.uk/elar/depositing-with-elar/who-can-deposit/). ELAR archives all materials from ELDP grantees (http://www.eldp.net), or materials where the majority of files are on open access.

The ASL-LEX Database (San Diego State, Tufts, and Boston Universities, http://asl-lex.org) is a database of lexical and phonological properties that have been compiled for nearly 1,000 signs of American Sign Language. The corpus is released free of use both for video streaming and downloading. The license enforced is the CC. The ASL-LEX database is available for download in .csv format from the Open Science Framework (https://osf.io/53vmf/). The data (clips and other information) have been stored through a custom repository which leverages on a commercial platform (YouTube). There is no indication that this repository is adequate for the purposes of the SIGN-HUB project.

f. European and International Platforms and Repositories

The above-mentioned foundational and critical core resources are continuously curating and capturing high-value reference datasets and fine-tuning them to enhance output, provide support for both human users and machine, and provide extensive tooling to access their content in rich, dynamic ways. However, not all datasets or even data types can be captured by, or submitted to, these repositories.

The accelerated proliferation of data from powerful new scientific instruments, simulations and the digitization of library resources has created a new impetus for increasing efforts and investments in order to tackle the specific challenges of data management and stewardship, and to ensure a coherent approach to research data access and preservation. Now we see the emergence of numerous general-purpose data repositories, at scales ranging from institutional (for example, a single university), to open globally-scoped repositories. In recent years, significant investments have been made by the European Commission and the European member states to create a pan-European e-infrastructure supporting multiple research communities.

As a result, a European e-infrastructure ecosystem has been established with communication networks, distributed computing and HPC facilities providing European researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. The most relevant examples are hereby listed and revised.

Dataverse (dataverse.org) is an open source web application to share, preserve, cite, explore and analyze research data. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive appropriate credit via a data citation with a persistent identifier. The Dataverse currently has multiple open APIs available, which allow for searching, depositing and accessing data. Dataverse is also installed in the countries of the European Union to preserve data collected by research communities of Netherlands, Germany, France and Finland. The largest Dataverse repository called DataverseNL and located in the Netherlands providing data management services for Dutch Universities. Among the academic installations, the most famous is the Harvard Dataverse (dataverse.harvard.edu), a repository for sharing, citing, analyzing, and preserving research data; open to all scientific data from all disciplines worldwide. CKAN (the Comprehensive Knowledge Archive Network, a web-based open source management system for the storage and distribution of open data, ckan.org) provides similar functions and is widely used for open data.

FigShare (figshare.com) is an online digital repository where researchers can preserve and share their research outputs, including figures, datasets, images, and videos. It is free to upload content and free to access, in adherence to the principle of open data. Figshare is a portfolio company of Digital Science, operated by Macmillan Publishers. Researchers can upload all of their research outputs to FigShare (up to a size constraint of 5GB), thus making them publicly available, with the aim to preview all of them in any desktop browser. Users can upload files in any format, and items are attributed a DOI. The current data types that can be chosen are figures, datasets, media (including video), papers (including pre-prints), posters, code, and filesets (groups of files). All files are released under a Creative Commons license, CC-BY for most files and CC0 (public domain) for datasets.

Dryad (datadryad.org) is a US disciplinary repository of data underlying scientific and medical publications. It makes data discoverable, freely reusable, and citable: its scientific, educational, and charitable mission is to promote the availability of data underlying findings in the scientific literature for research and educational reuse. The vision of Dryad is a scholarly communication system in which learned societies, publishers, institutions of research and education, funding bodies and other stakeholders collaboratively sustain and promote the preservation and reuse of data underlying the scholarly literature. Dryad aims to allow researchers to validate published findings, explore new analysis methodologies, re-purpose data for research questions unanticipated by the original authors, and perform synthetic studies such as formal

meta-analyses. Dryad serves as a repository for tables, spreadsheets, flat files, and all other kinds of published data for which specialized repositories do not already exist. Optimally, authors submit data to Dryad in conjunction with article publication, so that links to the data can be included in the published article. All data files in Dryad are associated with a published article, and are made available for reuse under the terms of a Creative Commons Zero waiver. The pricing planning of Dryad seems not very affordable for the project's need, asking to non-profit organization to pay 120$ for the first 20GB and 50$ for additional GB of digital content stored; nevertheless, its providers have been contacted to verify if the service meets the security requirements posed by the project itself (e.g., restrictions should apply on some of its data).

Zenodo (zenodo.org) is a service offered by OpenAIRE and the CERN; it is an open dependable home for the long-tail of science, enabling researchers to share and preserve any research outputs of any format and from any science, with a weight constraint for files up to 50GB; nevertheless, it is particularly suited for software and code, given its integration with GitHub to make code hosted in GitHub citable. It is released under the GPL license (version 2 or higher). Zenodo allow researchers to deposit both publications and data, while providing tools to link them.
The storage capacity of Zenodo seems not very affordable for the project's need, accepting datasets of maximum 50 GB of digital content stored; nevertheless, it is free, and its providers have been contacted to verify if the service meets the security requirements posed by the project itself (e.g., restrictions should apply on some of its data).

DataHub (datahub.csail.mit.edu) is a data ecosystem for individuals and trams managed by the MIT, licensed open source project from MIT CSAIL's Living Lab. It is a repository to store data centrally, without having to set up custom databases offering tools and software suites both to process data and to seamlessly share them with the academic community.

Mendeley Data (data.mendeley.com) is a secure cloud-based repository where researchers can store data, ensuring it is easy to share, access, and cite. Datasets can be shared privately amongst individuals, as well as published to share with the world. It offers facilities to quickly and easily upload files of any type; these files will have a permanent and referable home on the web, and will be always accessible via a unique and persistant DOI that is issued on publication. In particular, the citation is ensured to be Force11 and thus FAIR compliant. Data uploaded to Mendeley are stored on Amazon S3 servers, in Ireland, where it benefits from redundancy and multiple backups, and are released leveraging on a license to pick up from a wide range of Creative Commons and open software licenses. In addition, long-term data preservation is ensured by the partnership with DANS.

DANS (Data Archiving and Networked Services, www.dans.knaw.nl) is an industry-leading scientific data archive service, manabed by an institute of KNAW and NOW and representing the Netherlands institute for permanent access to digital research resources. To DANS it is possible to deposit open access research data to increase the visibility and findability of the data itself, which is maintained accessible in a sustainable form: in particular, this service is accessed through DataverseNL, a network of data repositories, which uses software developed by Harvard University. The platform is used worldwide. DataverseNL is jointly offered by participating institutes and DANS. DANS has been managing the network since 2014; the participating institutes are responsible for managing the deposited data in the local repositories. Institutes pay a fixed fee for participation in DataverseNL, to which storage costs for the data are added. Every participating institute has one vote in the Advisory Board, that determines the policy of DataverseNL. The Preservation Policy document explains the principles used by DANS in archiving digital research data in a sustainable form, including:

- the approach for sustainably archiving data, which partly goes back to the first datasets in the Steinmetz archive (established in 1964 and taken over by DANS in 2005);
- the discussion over data's authenticity;
- the archiving procedure, such as depositing data, sustainably and securely storing and making data available, and all of this in accordance with the international reference model for an Open Archival Information System.

The above-mentioned platforms and repositories accept a wide range of data types in a wide variety of formats: they generally do not attempt to integrate or harmonize the deposited data, and place few restrictions (or requirements) on the descriptors of the data deposition. The resulting data ecosystem, therefore, appears to be moving away from centralization, is becoming more diverse, and less integrated, thereby exacerbating the discovery and re-usability problem for both human and computational stakeholders.

For this reason, also platforms and repositories more specifically tailored to store, represent, convey, and preserve linguistic content have been in-depth analyzed. It is necessary to clarify whether these platforms support all the requirements posed by the SIGN-HUB project (e.g., for instance, whether they offer APIs to support the streaming of video content through the SIGN-HUB digital platform and if they offer support for authoritative mechanisms to let the general public dealing with restrict content) because they could host the project data, at least in part.

DARIAH (www.dariah.eu) is a pan-European networked infrastructure for arts and humanities scholars working with computational methods and technological tools. It supports digital research as well as the teaching of digital research methods. It connects dozens of research facilities over 17 European Countries. In addition, DARIAH has several cooperating partner institutions in countries not being a member of DARIAH, and strong ties to many research projects across Europe. Its mission is to provide services to scholars in the arts and humanities and therefore helping to do research at its best. DARIAH's vision is to enhance and support digitally-enabled research across the arts and humanities, and to facilitate the provision of services and activities for the digital arts and humanities research community in Europe and beyond. DARIAH integrates national digital arts and humanities initiatives in Europe and operates a platform to enable trans-national research. It offers a portfolio of services and activities centered around research communities and develops a research infrastructure for sharing and sustaining digital arts and humanities knowledge. By bringing together national activities from member countries, DARIAH will be able to offer a broad spectrum of services including training initiatives, such as summer schools and trans-national curricula, a knowledge repository with standards and good practices for digital asset management, and guidance on repository certification. Platforms for data sharing and digital publishing will be offered alongside technical systems for persistent identification, authentication, and long-term preservation.

CLARIN (Common Language Resources and Technology Infrastructure, www.clarin.eu) makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. It is a network grouping more than 40 centers doing research in linguistics and digital humanities (of which 18 are certified as B-centers, so compliant with the strongest requirements posed by the management of CLARIN in terms of FAIR-ness). It is a research infrastructure that was initiated in 2012 with the goal of building digital language resources and tools that are accessible through a federated identity with a single sign-on online environment for the support of researchers in the humanities and social sciences. Its mission is to create and maintain an infrastructure to support the sharing, use, and sustainability of language data and tools for research in the humanities and social sciences. CLARIN offers long-term solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area. Currently CLARIN provides easy and sustainable access to digital language data (in written, spoken, or multimodal form) for scholars in the social sciences and humanities, and beyond. CLARIN also offers advanced tools to discover, explore, exploit, annotate, analyze or combine such data sets, wherever they are located. This is enabled through a networked federation of centers: language data repositories, service centers and knowledge centers, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centers are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. One of the fundamental services of the CLARIN infrastructure is making sure that language resources can be archived and made available to the community in a reliable manner. To help researchers to store their resources (e.g., corpora, lexica, audio and video recordings, annotations, gram-

mars, etc.) in a sustainable way, many of the CLARIN centers offer a depositing service. They are willing to store the resources in their repository and assist with the technical and organizational details. This has a wide range of advantages:

- Long-term archiving: a storage guarantee can be given for a long period (up to 50 years in some cases);
- Resources can be cited easily with a persistent identifier;
- The resources and their metadata will be integrated into the infrastructure, making it possible to search them efficiently;
- Password-protected resources can be made available via an institutional login;
- Once resources are integrated in the CLARIN infrastructure, they can be analyzed and enriched more easily with various linguistic tools.

The technology infrastructure of CLARIN is ESFRI ERIC status since 2012, Landmark since 2016; basically, it provides easy and sustainable access for scholars in the humanities and social sciences and beyond to digital language data (in written, spoken, video or multimodal form) as well as advanced tools to discover, explore, exploit, annotate, analyze, or combine them, wherever they are located through a single sign-on online environment. This infrastructure is developed through a distributed architecture: http-accessible files, web applications, and web services are spread all over Europe.

EUDAT (eudat.eu), the European Data Infrastructure, was launched to target a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. EUDAT's mission is to design, develop, implement, and offer common data services to all interested researchers and research communities. These common data services obviously must be relevant to several communities, be available at European level, and they need to be characterized by a high degree of openness:

- Open Access should be the default principle;
- Independence from specific technologies should be guaranteed since these will change frequently;
- Flexibility to allow new communities to be integrated which is not a trivial requirement given the heterogeneity and fragmentation of the data landscape.

EUDAT is addressing these challenges by exploiting new technologies, integrating and building data-oriented services following the vision of a pan-European Collaborative Data Infrastructure. Its vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centers, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centers. Currently, EUDAT is working with more than 30 scientific communities and has built a suite of five integrated services – B2DROP, B2SHARE, B2SAFE, B2STAGE, and B2FIND – aiming at assisting them in resolving their grand challenges. The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Concerning the challenges posed by the management and distribution of digital content, especially video, the SIGN-HUB project has to deal with:

- ensuring long-term content persistence and easy access;
- data upload for subsequent access and interoperability (e.g., streaming purposes);
- implementing authentication and authorization mechanisms that are needed to restrict the access to some of the digital content;
- creation and management of linguistic-oriented metadata.

In this, the above-mentioned platforms seem suitable to host the project data, at least in part; in addition, building a custom FAIR-compliant repository ex-novo seems, given the present

state of the art, quite an unreasonable choice, which will lead to consistent expenses for the project and would drag resources (e.g., in terms of money and time) out from the development of the interfaces and tools that represent really the main outcomes of the technical Work Package WP3. However, none of them seems to adhere totally to the requirements and the challenges posed by the SIGN-HUB project. At least a responsible for all of them has been contacted to explain the needs for the SIGN-HUB project and to negotiate about providing all the services needed. The definitive statements about which platform to use, if any, and which services are to build custom within SIGN-HUB will be made before December the 6th 2017.

To use DARIAH, SIGN-HUB should have to negotiate with its managers about providing a streaming server and coping with linguistic-oriented metadata, services that do not seem to be exposed by the platform, as of today. In this, ISIDORE has been initialized in 2009 by the old research infrastructure Adonis and is today a platform and a search engine allowing the access to digital data of Humanities and Social Sciences. Open to all and especially to teachers, researchers, PhD students and students, it relies with enrichments on the principles of semantic web and provides access to data in open access. ISIDORE proposes more than five million of resources of the whole world and enrichments are available in 3 languages: French, English and Spanish.

On the other hand, the service offered by DARIAH for coping with Authentication and Authorization seems to fit, at least partially, the requirements posed by the project. The DARIAH Authentication and Authorization Infrastructure (DARIAH AAI) is based on SAML and Shibboleth in the European higher education identity inter-federation eduGAIN for implementing a Federated Single Sign-On. This means any Web application should integrate with a so-called SAML Service Provider (SP). The SP will protect the application, driving the log-in process and providing your application with attributes about the user who has logged in using a SAML Identity Provider (IdP) at another organization. Still, researchers and students in organizations that do not operate a federated Identity Provider can request a DARIAH *homeless* account using the DARIAH SelfService. DARIAH AAI is integrated in Higher Education Federations using the SAML standard.

To use EUDAT, SIGN-HUB should have to negotiate with its managers about providing a streaming server and authoritative mechanisms which are not a standard function of B2SHARE. B2SHARE is instead a is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store and share small-scale research data from diverse contexts, offering the following features:

- integrated with the EUDAT collaborative data infrastructure
- free upload and registration of stable research data
- data assigned a permanent identifier, which can be retraced to the data owner
- data owner defines access policy
- community-specific metadata extensions and user interfaces
- openly accessible and harvestable metadata
- representational state transfer application programming interface (REST API) for integration with community sites
- data integrity ensured by checksum during data ingest
- professionally managed storage service – no need to worry about hardware or network
- EUDAT user support
- monitoring of availability and use.

The alternative service offered by the EUDAT community, B2DROP, as a secure and trusted data exchange service for researchers and scientists to keep their research data synchronized and up-to-date and to exchange with other researchers, seems instead unfitted for the purposes of the project, given that through it users are offered only up to 20GB of storage space for research data.

On the other hand, the service offered by EUDAT for coping with Authentication and Authorization seems to fit, at least partially, the requirements posed by the project. B2ACCESS is an easy-to-use and secure Authentication and Authorization platform developed by EUDAT. B2ACCESS is versatile and can be integrated with any service. When B2ACCESS is integrated with a given service, the user may log in by using different methods of authentication:

- home organization identity provider;
- Google account;
- EUDAT ID.

EUDAT IDs are created by the B2ACCESS upon registration. Therefore, B2ACCESS is an Identity Providers for the users that do not have neither a Google account nor a Home Organization Identity Provider. In these cases, B2ACCESS offers also the tool for the managements of the EUDAT IDs. The features exposed by the B2ACCESS platform are:

- B2ACCESS supports several methods of authentication via the users' primary identity providers (OpenID, SAML, x.509)
- B2ACCESS can be used as primary identity provider, if necessary
- B2ACCESS can be integrated with any B2service and beyond
- B2ACCESS is integrated with EduGain and therefore support identities from theoretically hundreds of Universities and Research institutions around the world.
- B2ACCESS provides unique and persistent EUDAT IDs
- B2ACCESS allows group, community, and service managers to specify authorization decisions.

To use CLARIN, SIGN-HUB should have to negotiate with its managers about providing a streaming server and in particular authoritative mechanisms, as the solutions offered currently by CLARIN ERIC does not seem to fully adhere to the requirements of the SIGN-HUB project (e.g., federated login with research centers and higher education institutes is not an option when dealing with a general audience of broad end users that do not come from any scholar or research background). In this, the functionalities already exposed by CLARIN may be enlarged resorting to technologies for authentication, such as OAuth, an open standard for access delegation, commonly used as a way for Internet users to grant websites or applications access to their information on other websites but without giving them the passwords. Generally, OAuth provides to clients a *secure delegated access* to server resources on behalf of a resource owner. It specifies a process for resource owners to authorize third-party access to their server resources without sharing their credentials. Designed specifically to work with Hypertext Transfer Protocol (HTTP), OAuth essentially allows access tokens to be issued to third-party clients by an authorization server, with the approval of the resource owner. The third party then uses the access token to access the protected resources hosted by the resource server. OAuth is a service that is complementary to and distinct from OpenID. OAuth is also distinct from OATH, which is a reference architecture for authentication, not a standard for authorization. However, OAuth is directly related to OpenID Connect (OIDC) since OIDC is an authentication layer built on top of OAuth 2.0. OAuth is also distinct from XACML, which is an authorization policy standard. OAuth can be used in conjunction with XACML where OAuth is used for ownership consent and access delegation whereas XACML is used to define the authorization policies (e.g. managers can view documents in their region).
The technical pillars of CLARIN include:

- federated identity - letting users login to protected data and services with their own institutional login and password
- persistent identifiers - enabling sustainable citations of electronic resources
- sustainable repositories - digital archives where language resources can be stored, accessed and shared

- flexible metadata and concept definitions - to ensure semantic interoperability when describing language resources (see also CMDI in the following)
- content search - offering a search engine for a wide range of language resources + tools to explore and compare languages.

The services offered by CLARIN include a digital portal, a repository (for linguistic digital content and tools, which are described by metadata), web services and applications, resource inventory. Metadata are produced by the single centers: CLARIN central repository harvests them (typically every 2 or 3 days) and manages storing and indexing. Concerning video, CLARIN centers may pose individual limitations, but they generally seem to support .mpeg and .eaf (ELAN-based synchronized annotations with video) files. Having said that, resorting to CLARIN seems to be a better solution for the SIGN-HUB project, with respect to the other possible solutions such as DARIAH and EUDAT, in terms of data integration, interoperability, and persistence. We aim particularly to target CLARIN centers in Germany such as the Berlin-Brandenburg Academy of Sciences and Humanities, specialized in TEI encoding, and the Institut für Maschinelle Sprachverarbeitung, which has a strong expertise in linguistically annotated corpora, corpus tools, tools and methods for robust morphological and syntactic analysis of multiple languages.

CLARIN might be used either as a private repository, to store inaccessible data still ensuring their persistence over time, or as a public repository, where also data copies suitable for web-based access (e.g., low-quality copies of video content for streaming) are stored. The CLARIN ERIC network has already been contacted, and a dedicated workshop has been hosted with its technical director, dr. Dieter Van Uytvanck, to discuss the details of the integration of CLARIN resources within the SIGN-HUB project (e.g., pricing model) concerning the usage of CLARIN repositories. The costs of using the other solution have to be still evaluated, too.

g. Distribution of Digital Content

In the context of the SIGN-HUB project, issues of content distribution essentially relate to real-time distribution (streaming) of video content, but not only. The project is going to collect a wide amount of text document and images for describing linguistic content (e.g., images of phoneme).

Electronic document distribution is the process of distributing text documents and images electronically, making them easily accessible to end users via HTTP.

The possibility to retrieve document and images content through the web is of outmost importance for many scenarios, ranging from commercial opportunities to the spread of knowledge and culture. Digital technology has the potential to improve access to research material, allowing access to precisely the content sought by an end user. This implies full content search and retrieval, so that users can get to precisely the page they are interested in for text, or precisely the sound or video clip for audio or video productions. To realize this potential, the content must be both described, so that its production attributes are preserved and so users can navigate to the content meeting their needs, and properly represented and delivered to the users.

Concerning the needs of the SIGN-HUB project, after having analyzed the requirements it poses, we may conclude that the project should deal with documents and images as well as it will deal with videos: a repository should be identified that accepts also documents and images for their storage and long-term preservation, which in addition offers capabilities for being linked to the SIGN-HUB digital platform, to be accessed via any web browser (also from tablet and mobile devices) from authorized users. Metadata should be enforced also to convey searchable information about the documents and images, to maximize their usability, findability, and interoperability.

h. Real-time Streaming of Video Content

Within the SIGN-HUB project, it will be required to produce, manage, and distribute two classes of digital video content, as follows:

- short clips (e.g., few seconds or minutes): these will be on one side answers and questions within the survey tool and short videos of the signers participating in the T2.3 pilots and will have to be stored directly within, and managed by, the SIGN-HUB digital platform, since it is sensitive data, on the other side there will be also short video clips for the Grammarians (Working Package WP2.1), which might be stored in an external repository or platform but still be used to be showed in the platform;
- long clips (e.g., one hour or more): these clips will constitute the digital archive of old signers' linguistic and cultural heritage and should be preserved for long time.

So, SIGN-HUB will have to describe, annotate, and stream also very long videos. Considering this, it is recommended to identify a proper streaming service.

On the one hand, we are still verifying whether existing research platforms and repositories, such as CLARIN, offer a streaming service suitable for the needs of the project (i.e., if they offer APIs to expose in streaming a video stored in their repository towards the SIGN-HUB digital platform). We have, instead, already discarded the option to refer to commercial streaming providers, such as YouTube or Video, as they are not compatible with the requirements posed by deliverable D1.5 (i.e., our Data Management Plan). So, it is still possible that within SIGN-HUB we will have to develop a custom streaming service: this is why this section deals with technologies and protocols for real-time streaming of digital video content.

On the other hand, it would be nice to identify, or even develop, a streaming service that may deal with video fragmentation, allowing also to present and transmit short video fragments (e.g., even just a sentence or an expression for a certain annotation) instead of the whole video, which saves network and streaming capacity. ELAN video player supports the fragment option of online video streaming, which has been used in the DOBES project. However, although being a nice and useful option, this is not a strict requirement for the SIGN-HUB project.

According to Cisco[9], video services are fast becoming an essential part of consumers' lives. In fact, in a recent white paper about global mobile data traffic the company foresees that mobile video will increase 11-fold between 2015 and 2020, accounting for 75 percent of total mobile data traffic by the end of 2020. Similar percentages and traffic will be registered for non-mobile platforms, especially when smart and high-definition television is considered. As in the case of mobile networks, video devices can have a multiplier effect on traffic. An Internet-enabled television that draws 45 minutes of high-definition video content per day from the Internet can generate as much Internet traffic as an entire household today. With the growth of video viewing on smartphones and tablets, traffic from these devices is growing as a percentage of total Internet traffic. These numbers are almost self-explanatory, and describe why researchers have, in the last years, tried to define standards for representing videos that would require the least possible amount of resources offering at the same time the highest possible quality of service, and to build adequate infrastructures (in terms of hardware, software and protocols) to deliver video content to customers.

As analog video collections are digitized and new video is created in digital form, users will have unprecedented access to video material, getting what they need, when they need it, wherever they happen to be. Such a vision assumes that video can be adequately stored and distributed with appropriate rights management, as well as indexed to facilitate effective information retrieval. So, it comes again to an argument that has been already dealt with previ-

[9] *"Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update , 2010 – 2015," Growth Lakel., vol. 2011, no. 4, pp. 2010–2015, 2011.*

ously: metadata for video are crucial when one considers the huge volume of bits within digital video representations.

Metadata will continue to document the rights of producers and access controls for consumers. Combined with electronic access, metadata enable remuneration for each viewing or performance down to the level of individual video segments or frames, rather than of distributions or broadcasts. Metadata can grow to include specific usage information; for example, which portions of the video are played, how often, and by what sorts of users in terms of age, sex, nationality, and other attributes. Of course, such usage data should respect a user's privacy and be controlled through optional inclusion and specific individual anonymity.

Metadata provide the window of access into a digital video archive. Without metadata, the archive could have the perfect storage strategy and would still be meaningless, because there would be no retrieval and hence no need to store the bits. With appropriate metadata, the archive becomes accessible. By enhancing the metadata, the archive can remain fresh and current and accessible efficiently and effectively; metadata can be updated without changing video content: only the metadata are enhanced, which in turn enhances the value of the video archive[10].

Streaming of Large Data

Due to the increasing demand for multimedia information on the web, streaming video over the Internet has received tremendous attention from academia and industry. Transmission of real-time video typically has strict quality requirements in terms of bandwidth, delay, and data loss. However, current Internet network is based on the best-effort principle, meaning that it does its best being unable to guarantee any quality of service to streaming video. Furthermore, for video multicast, it is difficult to achieve both efficiency and flexibility. Thus, Internet streaming video poses many challenges.

To address these challenges, extensive research has been conducted in the last years. Recent advances in computing technology, compression technology, high-bandwidth storage devices, and high-speed networks have made it feasible to provide real-time multimedia services over the Internet. Real-time multimedia, as the name implies, has very strict timing constraints. For example, audio and video data must be played out continuously. If the data does not arrive in time, the playout process will pause, which is annoying to human ears and eyes. Real-time transport of live video or stored video is the predominant part of real-time multimedia.

Storing hundreds of billions (and even billions of billions) of Bytes of data is no longer uncommon for organizations throughout a vast array of industries, but there's a big difference between big data sets composed of stagnant archived files and big data containing streaming media files and video contents. These are commonly referred to as big data at rest, and big data in motion. The implications of video growth are difficult to overstate. With video growth, Internet traffic is evolving from a relatively steady stream of traffic (characteristic of peer-to-peer traffic) to a more dynamic traffic pattern.

Apart from short-form video and video calling, most forms of Internet video do not have a large upstream component. As a result, traffic is not becoming more symmetric, a situation that many expected when user-generated content first became popular. The emergence of subscribers as content producers is an extremely important social, economic, and cultural phenomenon, but subscribers still consume far more video than they produce. Upstream traffic has been slightly declining as a percentage for several years.

[10] *H. D. Wactlar and M. G. Christel, "Digital video archives: Managing through metadata," Build. a Natl. Strateg. Digit. Preserv. Issues Digit. media Arch., vol. 84, p. 99, 2002.*

It appears likely that residential Internet traffic will remain asymmetric for the next few years. However, numerous scenarios could result in a move toward increased symmetry. For instance, content providers and distributors could adopt peer-to-peer distribution mechanisms. Generally, if service providers provide ample upstream bandwidth, applications that use upstream capacity will begin to appear[11].

Peer-to-Peer (P2P) networking has recently emerged as a new paradigm to build distributed network applications. The basic design philosophy of P2P is to encourage users to act as both clients and servers, namely as peers. In a P2P network, a peer not only downloads data from the network, but also uploads the downloaded data to other users in the network. The uploading bandwidth of end users is efficiently utilized to reduce the bandwidth burdens otherwise placed on the servers.

Video streaming can be classified into two categories: live and on-demand. In a live streaming session, a live video content is disseminated to all users in real-time. The video playbacks on all users are synchronized. On the other hand, video-on-demand users enjoy the flexibility of watching whatever video clips whenever they want. Playbacks of the same video clip on different users are not synchronized[12].

To provide quality multimedia presentations, adequate support from the network is critical. This is because network support can reduce transport delay and packet loss ratio. Streaming video and audio are classified as continuous media because they consist of a sequence of audio samples and video frames that convey meaningful information only when presented in time. Built on top of the Internet, continuous media distribution services are designed with the aim of providing the required quality of service and achieving efficiency for streaming video/audio over the best-effort Internet.

Video streaming is an important component of many Internet multimedia applications, such as distance learning, digital libraries, home shopping, and video-on-demand. The best-effort nature of the current Internet poses many challenges to the design of streaming video systems. Network support is important to provide quality multimedia presentations. Continuous media distribution services are built on top of the best-effort Internet with the aim of ensuring the necessary quality-of-service and efficiency for streaming video. A major topic of active research is how to build a scalable, efficient, cost-effective and incremental deployable infrastructure for continuous media distribution[13].

Media Streaming Communications Protocols

Quite a few protocols have been designed and standardized for communication between clients and streaming servers. According to their functionalities, the protocols directly related to Internet streaming video can be classified into the following three categories.

Network-layer protocols provide basic network service support such as network addressing. The Internet Protocol (IP) serves as the network-layer protocol for Internet video streaming.

Transport protocols provide end-to-end network transport functions for streaming applications. Transport protocols include User Datagram Protocol (UDP), Transmission Control Protocol (TCP), real-time transport protocol (RTP), and real-time control protocol (RTCP). UDP and TCP are lower-layer transport protocols while RTP and RTCP are upper-layer transport protocols,

[11] *CISCO, "The Zettabyte Era: Trends and Analysis," Cisco, no. May 2015, pp. 1–29, 2015.*

[12] *Y. Liu, Y. Guo, and C. Liang, "A survey on peer-to-peer video streaming systems," Peer-to-Peer Netw. Appl., vol. 1, no. 1, pp. 18–28, 2008.*

[13] *D. Wu, Y. T. Hou, W. Zhu, Y. Q. Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions," IEEE Trans. Circuits Syst. Video Technol., vol. 11, no. 3, pp. 282–300, 2001.*

which are implemented on top of UDP or TCP, these last protocols implementing functions as multiplexing, error control, congestion control, or flow control. HTTP Live Streaming is another protocol lying in this category.

Session control protocol define the messages and procedures to control the delivery of the multimedia data during an established session. Examples are the RTSP and the session initiation protocol (SIP).

i. HTTP Live Streaming

HTTP Live Streaming (also known as HLS) is an HTTP-based media streaming communications protocol implemented by Apple Inc. as part of its QuickTime, Safari, OS X, and iOS software. It is like MPEG-DASH in that it works by breaking the overall stream into a sequence of small HTTP-based file downloads, each download loading one short chunk of an overall potentially unbounded transport stream. As the stream is played, the client may select from several different alternate streams containing the same material encoded at a variety of data rates, allowing the streaming session to adapt to the available data rate. At the start of the streaming session, it downloads an extended playlist (using the M3U data format) containing the metadata for the various sub-streams which are available.

Since its requests use only standard HTTP transactions, HTTP Live Streaming can traverse any firewall or proxy server that lets through standard HTTP traffic, unlike UDP-based protocols such as RTP. This also allows content to be offered from conventional HTTP servers as origin and delivered over widely available HTTP-based content delivery networks.

HTTP Live Streaming is flexible, allowing high playback quality in wireless networks with high bandwidth and low quality when the bandwidth is reduced. HTTP Live Streaming also provides protection against errors, generating alternative different flows video to use them if there are any errors in segment. To make the system scalable and adaptable to the bandwidth of the network, the video flow is coded in different qualities. Thus, depending on the bandwidth and transfer network speed, the video will play at different qualities. To implement this, the system must encode the video in different qualities and generate an index file that contains the locations of the different quality levels.

The client software internally manages the different qualities, making requests to the highest possible quality within the bandwidth of the network. One issue that can arise by relying on the client-side system is that a user may experience different bitrates throughout the duration of playback. This can be avoided by delaying the responsiveness of the player to changes in client-side bandwidth.

ii. Real-time Transport Protocol (RTP)

RTP is an Internet standard protocol designed to provide end-to-end transport functions for supporting real-time applications. Real-time Control Protocol, RTCP, is a companion protocol designed to work in conjunction with RTP; it is designed to provide quality-of-service feedback to the participants of an RTP session. Simplifying the concept, RTP is a data transfer protocol while RTCP is a control protocol. RTP does not guarantee quality-of-service nor reliable delivery, but rather, provides some interesting functions in support of media streaming, such as time-stamping and sequence numbering (RTP employs sequence numbering to place the incoming RTP packets in the correct order, since in-order delivery is not guaranteed from the Internet network; the sequence number is also used for packet loss detection), identification of payload type and source.

In an RTP session, participants periodically send RTCP packets to convey feedback on quality of data delivery and information of membership. Basically, RTCP provides feedback to an application regarding the quality of data distribution in form reports sent by the source and by the receiver. The reports can contain information on the quality of reception such as fraction and cumulative number of lost RTP packets since the last report.

i. Considerations

Having in-depth considered the requirements posed by the SIGN-HUB project, as exposed by deliverable D3.1, the outcomes of our *build or buy* analysis say that for the purposes of the SIGN-HUB project it is recommendable to refer directly to large repositories and platforms, possibly deriving from other research projects in the area of digitalizing humanities content, that embed capabilities for video management and distribution, and ensure long-term persistence of digital content, both video or not, and compliance to the FAIR principles. This document identifies three main platforms (i.e., CLARIN, DARIAH, and EUDAT). What exactly the integration with external platforms will cost to the SIGN-HUB project is not clear to date.

If the costs will be judged as too high compared to the benefits exposed, we will have to build a SIGN-HUB custom repository: in this, we will surely leverage on Islandora and Fedora Commons, which expose features that are of interest for the SIGN-HUB project. Then, we will have to add mechanisms to support for identification (through CLARIN PID, if this will be judged feasible, or DOI) and description (through CLARIN CMDI, if this will be judged feasible, or IMDI) of digital content.

Our analysis revises a wide range of video CMSs, both proprietary and open source. These solutions seem to be either expensive either lacking of codecs that are instead required to convert videos for streaming purposes. It is also not clear how they threat intellectual propriety of the digital content and which services (e.g., APIs) they may expose to allow an easy integration with any digital platform (e.g., to enable video streaming).

3. Management and Distribution of Surveys

The SIGN-HUB digital platform will comprise an interface for creating and delivering surveys, to manage on-line surveys that use stimuli with a variety of format (text, picture, and video). In this, capabilities for managing multimedia content, and in particular videos, will be of paramount importance. This chapter deals with existing technologies and platforms that represent the state of the art for web-based tools to create, delivery, and manage surveys.

In the following, a list of technologies for administrating surveys is given.

Formstack (www.formstack.com) is a data management solution that helps users collect information through various types of online forms, including surveys, job applications, event registrations, and payment forms. It includes several add-ons and paid features (e.g., integration with workflows, with PayPal). Formstack enables organizations to build and design web forms without any programming or software skills required. Formstack users can embed mobile-ready forms on their websites and social media profiles, collect online payments, gather feedback from customers and employees, and create process workflows for their organization. It provides users with a friendly front-end where surveys can be built by simply drag-and-drop and customization of basic elements and outcomes can be filtered and exported easily in a variety of format (e.g., CSV, XLS, RTF, PDP). Surveys created with Formstack are typically fully accessible and compliant with the US regulations (Section 50814).

Qualtrics (www.qualtrics.com) provides a web-based software that enables users to do many kinds of online data collection and analysis including market research, customer satisfaction and loyalty, product and concept testing, employee evaluations and website feedback. It is often used in academic studies. It delivers similar functionalities w.r.t. Formstack, adding on top dedicated modules for powerful analytics and statistics on data collected from the surveys.

Google Form (www.google.com/forms/about) is a well-known and free tool that enables users to create and manage simple forms, typically comprising few sections and questions requiring users mostly basic interactions. The information collected from each survey is automatically connected to a spreadsheet. The spreadsheet is populated with the survey and quiz responses. Google Form features include, but are not limited to, menu search, shuffle of questions for randomized order, limiting responses to once per person, shorter URLs, custom themes, automatically generating answer suggestions when creating forms, and an "Upload file" option for users answering to share content through.

Typeform (www.typeform.com) is an online software as a service (SaaS) company that specializes in online form building. Its main software creates dynamic forms based on user needs. The "typeforms" present questions which slide down one after another showing only one question at a time to keep users engaged and can include images, and GIFs or videos. The tool includes "Calculator," custom "Thank You" screens, "Question Groups" which allow questions to be added to sections or include sub-questions and "Logic Jump" which customizes the questions a user sees based on their selections. The form can be embedded into a website, open in a pop-up, or be accessed through a unique URL. The form-builder uses a freemium business model. Surveys created with Typeform are typically very easy to follow and be taken even for not-expert users.

Formsite (www.formsite.com) is a service that enables non-technical users to build professional and responsive web forms and surveys with no HTML or coding experience. Formsite forms offer built-in validation and error handling, as well as the ability to process, store, and email form submissions. Like many software as a service business models, features available to Formsite users are based on an annual subscription fee, which varies depending on the level of service that the user desires, an ad supported, free level of service is offered for light users,

[14] *https://www.epa.gov/accessibility/what-section-508*

along with more expensive options that allow heavier users to create larger forms and store more results. Formsite seems to lack any support for having video content inserted in the surveys, either for the questions either for the answers. It integrates seamlessly with payment capabilities (e.g., PayPal) and third-party services to share the outcomes of the surveys (e.g., Google Drive, Dropbox).

123contactform (123contactform.com) is a free online service that allows to easily create and manage surveys via a custom drag-and-drop interface. Its main functionalities are similar to the ones exposed by the other online services, additionally however supporting SMS notifications, multi-language forms, and integration with mailing list services (e.g., MailChimp).

All of these solutions, either being free or requiring some fee to be used, provide interesting features. Formstack ensures to achieve professionally-looking outcomes; in addition, it produces fully accessible forms (i.e., forms that are 508 compliant, meaning that can be accessed by all users, regardless of their disability status) and databases. Google Forms, despite being well-known, cannot easily scale up when the database size increases and it is therefore difficult to access and manage.

Having in-depth considered the requirements posed by the SIGN-HUB project, the outcomes of our *build or buy* analysis say that all the above-mentioned software and services for the management and distribution of services must be discarded, because they are generally too complex, and incapable of coping neither with even brief videos to be uploaded as answers to questions nor with fonts specific for handshapes, and require data to be hosted on the cloud (while short videos for the surveys should be hosted on premises identified by the SIGN-HUB project). Thus, no tool is suitable as it is for the need of the SIGN-HUB digital platform: nevertheless, the development of its capabilities concerning surveys should be inspired by what we have seen for these tools (e.g., modalities to export of data, building of the survey by drag-and-drop of basic building blocks).

4. Standards for Archiving and Preservation of Digital Content

a. Introduction

Multimedia content within web platforms and systems has to be organized with particular care. In fact, multimedia content is different from regular text document or articles because it is more difficult to store (since it tends to weight more in terms of space) and to be automatically processed and analyzed (since it is difficult for computers), even though CMSs like the ones above described expose functionalities and tools to ease these tasks. For these reasons, the task of guaranteeing a simple, standard and above all persistent access to information stored within multimedia content is not trivial.

Digital preservation can be understood as the range of activities required to ensure that digital objects remain accessible for as long as they are needed. In a popular definition, Hedstrom says that digital preservation involves "the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable". Despite the growing ubiquity of digital information, the long-term preservation of information in digital form is far from a simple task. According to the Harrod's Librarian Glossary, digital preservation is the method of keeping digital material alive so that they remain usable as technological advances render original hardware and software specification obsolete. Digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to digital content over time[15].

To standardize digital preservation practice and to provide a set of recommendations for preservation program implementation, the Reference Model for an Open Archival Information System (OAIS) was developed. OAIS deals with various aspects of a digital object's life cycle: ingest, archival storage, data management, administration, access and preservation planning. The model also addresses metadata issues and recommends that five types of metadata be attached to a digital object: reference (identification) information, provenance (including preservation history), context, fixity (authenticity indicators), and representation[16].

In March 2000, the Research Libraries Group (RLG) and Online Computer Library Center (OCLC) began a collaboration to establish attributes of a digital repository for research organizations, building on and incorporating the emerging international standard of the Reference Model for an OAIS. In 2002, they published "Trusted Digital Repositories: Attributes and Responsibilities". In that document, a "Trusted Digital Repository" (TDR) is defined as "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future". A TDR must include attributes such as compliance with the reference model for an OAIS, administrative responsibility, organizational viability, financial sustainability, technological and procedural suitability, system security, procedural accountability. The report also recommended the collaborative development of digital repository certifications, models for cooperative networks, and sharing of research and information on digital preservation with regard to intellectual property rights[17].

---

[15] *Digital Preservation Coalition, "Introduction: Definitions and Concepts," Digital Preservation Handbook, 2008, available on-line at http://handbook.dpconline.org/ (accessed on 04 July 2016)*

[16] *Digital Preservation Coalition, "Digital Preservation Coalition Interactive Assessment: Selection of Digital Materials for Long-term Retention" 2006, available on-line at http://dpconline.org/advice/preservationhandbook/decision-tree/decision-tree-interactive-assessment (accessed on 04 July 2016)*

[17] *RLG-OCLC, "Trusted Digital Repositories: Attributes and Responsibilities," Mountain View, CA, 2002*

While digital multimedia content is typically easier to create and to keep up-to-date, at the same time there are many challenges in its preservation of this content. In fact, digital object always needs a software environment to render it (e.g., a CMS). These environments keep evolving and changing at a rapid pace, threatening the continuity of access to the content. At the heart of the problem is the rapid obsolescence of the various technologies on which digital information depends. Physical storage media, data formats, hardware, and software all become obsolete over time, posing significant threats to the survival of the content. Rapidly changing technologies can hinder digital preservationists work and techniques due to outdated and antiquated machines or technology. The economic challenges of digital preservation are also great. Preservation programs require significant upfront investment to create, along with ongoing costs for data ingest, data management, data storage, and staffing. One of the key strategic challenges to such programs is the fact that, while they require significant current and ongoing funding, their benefits accrue largely to future generations.

Hence, for long-term preservation, digital video presents a number of challenges. Regardless of how these challenges are addressed, modern digital technology has huge size, but also huge potential, for facilitating access to video archive material.

Digital technology has the potential to improve access to the desired multimedia content. This implies full content search and retrieval, and it is typically done relying on content filename and metadata indexing. Creating such metadata by hand is prohibitively expensive and inappropriate for digital video, where much of the metadata is a by-product of the way in which the artifact is generated. Current research is focusing on developing reliable automated techniques for metadata creation[18].

The key to the successful implementation of all preservation strategies will be the capture, creation, maintenance and application of appropriate metadata. Understood in this way, it is clear that such metadata needs to support an extremely wide range of different functions, including discovery and access, recording the contexts and provenance of objects, to the documentation of repository actions and policies. Conceptually, therefore, preservation metadata spans the traditional division of metadata into descriptive, structural and administrative categories. Within digital repositories, metadata should accompany and make reference to digital objects, providing associated descriptive, structural, administrative, rights management, and other kinds of information. The wide range of functions that preservation metadata is expected to support means that the definition (or recommendation) of standards is not a simple task. The situation is complicated further by the knowledge that different kinds of metadata will be required to support different digital preservation strategies and that the metadata standards themselves will need to evolve over time. To date, the information model defined by the OAIS Reference Model has been extremely influential on the development of preservation metadata standards. Nevertheless, there exist additional requirements for metadata when data preservation must be ensured.

Most of the existing digital repositories are today based on this model (although often greatly simplifying the original structure of information components and features), and also on this model have leveraged many international research projects struggling to define shared schemes and procedures for the standardization of long-term data preservation, such as the Online Computer Library Center/Research Libraries Group (OCLC/RLG) Metadata Framework (2002) and the subsequent PREMIS, with the aim of developing a set of crucial and easily implementable object preservation elements digital and, more generally, digital documentary systems.

PREMIS (Preservation Metadata Implementation Strategies, www.loc.gov/standards/premis/v2/premis-2-0.pdf) was born in 2003 as the product of a workgroup set by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG), which included more than 30 representatives from all over the world, with the

[18] *M. Day, "The long-term preservation of Web content," in Web archiving, Springer, 2006, pp. 177–199*

mandate to define a basic set of metadata that it could be concretely usable, complete with guidelines for management and use. The result of the working group was published in 2005 under the heading Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. Subsequently, a revised version (2.0) was released in 2008, which is still the main reference for metadata for data preservation. The main findings of the working group consist of:

- an essential set of metadata consistent with other relevant standards of descriptive metadata based on specific domains) predisposed in the form of an XML schema;
- a data dictionary or data dictionary designed to facilitate the use of the elaborated scheme.

The main components of the final product (schema and data dictionary) were completed and approved in May 2005 and concern (consistent with the requirements of ISO 14721 OAIS) about the metadata for preservation of digital content, that is, the information necessary to ensure the availability, accessibility, intelligibility and authenticity of digital resources. Particular care has been paid to documenting data origin (the object's history) and the relations between different objects (especially within the digital deposit). PREMIS focuses almost exclusively on describing and treating objects and events, considering that a digital deposit has to do with objects to be preserved and with events that interact with objects in conservative processes, and that the definition and description of intellectual entities and agents should be deepened by the experts of each domain in relation to their industry standards

In addition to the PREMIS standard, unanimously recognized as the main reference for conservation metadata, other standards are also relevant for the present analysis, although being developed in other areas, such as:

- ISO 23081-1/3: 2011 - Metadata for records, a group of ISO metadata for document management that has developed a general model of metadata articulated into functional document management categories (document metadata, policy metadata, metadata on producer, process and document management processes). Such standard, published in three parts between 2006 and 2011, regulates the processing and use of metadata required for the management of administrative documents (records) set up in accordance with ISO 15489. The standard defines which metadata are needed to achieve the goals set by ISO 15489 (including authenticity, security, integrity, and usability). It does not include any convergence between the metadata required for records management and archive description rules, but emphasizes the importance of migrating metadata from records management systems to archive systems. However, this standard is not available freely (only a small preview is actually accessible).
- NLZ - National Library of New Zealand - Metadata Implementation Schema, a fairly comprehensive proposal developed by the National Library of New Zealand to manage the conservation of its collections, specifies the elementary data needed to support the storage of digital objects and forms the basis for project repository and input systems for collecting and storing metadata.

This metadata schema is associated with a Metadata Extraction Tool designed for:

- Automatically extract metadata for storage from digital files;
- Produce an output in XML format, containing metadata and usable in conservation activities.

The tool has been specifically tailored to target tasks related to preservation of digital content, but can also be configured to represent metadata in other contexts.

b. Identification of resources

A persistent identifier is a long-lasting reference to a digital resource. Typically, it is made up of two components: a unique identifier and a service that locates the resource over time, even when its physical location changes. The first component helps to ensure the provenance of a digital resource, while the second one ensures that the identifier is resolved to the correct current location. Persistent identifiers thus aim to solve the problem of the persistence of the digital content. A simple web address (links) is often not enough, since it may fail to take the user to the referenced resource expected (e.g., for technical problem or even simply because the resource has been physically moved from its original location). Problems in accessing to resources are frustrating for every user, but particularly for researchers have a direct and considerable impact. Persistent identifiers can also be used behind-the-scenes within a repository to manage some of the challenges in cataloguing and describing, or providing intellectual control and access to born-digital materials.

Since the problem of persistence of an identifier is created by humans, the solution of persistent identifiers also has to involve people and services not just technologies. There are several persistent identifier schemes and all require a human service element to maintain their resolution systems. The main persistent identifier schemes currently in use are detailed below.

DOIs are digital identifiers for objects (whether digital, physical or abstract) which can be assigned by organizations in membership of one of the DOI Registration Agencies; the two best known ones are CrossRef, for journal articles and some other scholarly publications, and DataCite for a wide range of data objects. As well as the object identifier, DOI has a system infrastructure to ensure a URL resolves to the correct location for that object.

Handles are unique and persistent identifiers for Internet resources, with a central registry to resolve URLs to the current location. Each Handle identifies a single resource, and the organization which created or now maintains the resource. The Handle system also underpins the technical infrastructure of DOIs, which are a special type of Handles.

ARK (Archival Resource Key) is an identifier scheme conceived by the California Digital Library (CDL), aiming to identify objects in a persistent way. The scheme was designed on the basis that persistence "is purely a matter of service and is neither inherent in an object nor conferred on it by a particular naming syntax".

PURLs (Persistent Uniform Resource Locators) are URLs which redirect to the location of the requested web resource using standard HTTP status codes. A PURL is thus a permanent web address which contains the command to redirect to another page, one which can change over time.

URNs (Universal Resource Names) are persistent, location-independent identifiers, allowing the simple mapping of namespaces into a single URN namespace. The existence of such a Uniform Resource Identifier does not imply availability of the identified resource, but such URIs are required to remain globally unique and persistent, even when the resource ceases to exist or becomes unavailable. The URN term is now deprecated except in the very narrow sense of a formal namespace for expressing a Uniform Resource Identifier.

There needs to be a social contract to maintain the persistence of the resolution service-either by the organization hosting the digital resource, a trusted third party or a combination of the two. Each scheme has its own advantages and constraints but it is worth considering the following issues when deciding on a persistent identifier strategy or approach for the SIGN-HUB project:

- there is no single system accepted by all, though DOIs are very well established and widely deployed;
- there may be costs to establishing or using a resolver service;
- there may be dependences on ongoing maintenance of the permanent identifier system.

Still, the advantages brought by persistent identifiers are too big (e.g., they are critically important in helping to establish the authenticity of a resource, they provide access to a resource even if its location changes, they allow interoperability between collections) that their usage has been in-depth analyzed and largely discussed within the SIGN-HUB project.

Clearly, DOI seems to be a valid option. In fact, the DOI system offers persistent, semantically-interoperable resolution to related current data and is best suited to material that will be used in services outside the direct control of the issuing assigner (e.g., public citation or managing content of value). It uses a managed registry (providing social and technical infrastructure). It does not assume any specific business model for the provision of identifiers or services and enables other existing services to link to it in defined ways. Several approaches for making identifiers persistent have been proposed. The comparison of persistent identifier approaches is difficult because they are not all doing the same thing. Imprecisely referring to a set of schemes as "identifiers" doesn't mean that they can be compared easily. Other "identifier systems" may be enabling technologies with low barriers to entry, providing an easy to use labeling mechanism that allows anyone to set up a new instance (examples include Persistent Uniform Resource Locator (PURL), URLs, Globally Unique Identifiers (GUIDs), etc.), but may lack some of the functionality of a registry-controlled scheme and will usually lack accompanying metadata in a controlled scheme. The DOI system does not have this approach and should not be compared directly to such identifier schemes. Various applications using such enabling technologies with added features have been devised that meet some of the features offered by the DOI system for specific sectors (e.g., ARK). A DOI name does not depend on the object's location and, in this way, is similar to a Uniform Resource Name (URN) or PURL but differs from an ordinary URL. URLs are often used as substitute identifiers for documents on the Internet (better characterized as Uniform Resource Identifiers) although the same document at two different locations has two URLs. By contrast, persistent identifiers such as DOI names identify objects as first class entities: two instances of the same object would have the same DOI name.

However, despite its abilities, the DOI system has drawn criticism from researchers for directing users to non-free copies of documents that would have been available for no additional fee from alternative locations. In addition, it is not directly tailored for identifying digital linguistic resources. So, we have evaluated the exploitation of identifiers specifically tailored for digital linguistic resources has been carefully considered. In this, CLARIN persistent identifiers represent the standard de facto and are in the following briefly revised.

The use of handles is deeply integrated into the CLARIN infrastructure. They are a requirement for B-centers (i.e., centers in CLARIN that are certificated as properly adhering to CLARIN principles and commitment) and form a part of the Language Resource inventory and the Virtual Collection Registry. CLARIN centers can acquire handles via the ERIC consortium, a collaboration between several major computing centers to offer stable and reliable persistent identifier services. Persistent identifiers in CLARIN are used from references between metadata descriptions and their resources up to references between semantic assertions made by using the RDF (Resource Description Framework). Handling identifiers in CLARIN is very simple: there exists a service that, starting from some required information on the data resource (in particular the path to access the resource such as a URL) provides directly the identifier (PID), which then can (and actually should) be inserted as a reference in the metadata of the data itself. PIDs are valuable instruments to guarantee long term preservation and accessibility. CLARIN has an arrangement with the ERIC consortium that CLARIN members will be able to both register and resolve PIDs. This consortium groups a number of reliable European service providers that want to participate in providing a redundant service for the research world. The service is based on the Handle System which according to our investigations is the only robust system meeting all requirements.

The CLARIN ERIC network has already been contacted, and a dedicated workshop has been hosted with its technical director, dr. Dieter Van Uytvanck, to discuss the details of the integration of CLARIN resources within the SIGN-HUB project (e.g., pricing model) concerning the usage of CLARIN PIDs. If it is possible, we would recommend for the SIGN-HUB project to rely on the usage of CLARIN PIDs and to integrate them within SIGN-HUB digital platform; otherwise, a custom solution based on DOI should be implemented. This choice and the final decision

should be taken mainly on the basis of a cost analysis, since both the possibility seems technical feasible and would bring benefit to the SIGN-HUB project.

c. Metadata

Metadata is often referred to as *data about data*. A metadata standard is a requirement which is intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users. Metadata is usually categorized in three types:

- Descriptive metadata describes an information resource for identification and retrieval through elements such as title, author, and abstract;
- Structural metadata documents relationships within and among objects through elements such as links to other components (e.g., how pages are put together to form chapters);
- Administrative metadata helps to manage information resources through elements such as version number, archiving date, and other technical information for purposes of file management, rights management and preservation.

Metadata elements grouped into sets designed for a specific purpose, e.g., for a specific domain or a particular type of information resource, are called metadata schemes. For every element, the name and the semantics (i.e., its meaning) are specified. Content rules (how content must be formulated), representation rules (e.g., capitalization rules), and allowed element values (e.g., from a controlled vocabulary) can be specified optionally. Some schemes also specify in which syntax the elements must be encoded, in contrast to syntax independent schemes. Many current schemes use Standard Generalized Markup Language (SGML) or XML to specify their syntax. Metadata schemes that are developed and maintained by standard organizations or organizations that have taken on such responsibility are called metadata standards. The most relevant standards for the purposes of this analysis are briefly listed in the following.

Dublin Core (www.dublincore.org) is a small set of vocabulary terms that can be used to describe web resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks. The full set of Dublin Core metadata terms can be found on the Dublin Core Metadata Initiative (DCMI) website. The original set of 15 classic metadata terms, known as the Dublin Core Metadata Element Set, is endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-2009
- NISO Standard Z39.85.

Dublin Core Metadata may be used for multiple purposes, from simple resource description, to combining metadata vocabularies of different metadata standards, to providing interoperability for metadata vocabularies in the Linked Data cloud and Semantic Web implementations. Starting in 2000, the Dublin Core community focused on "application profiles" – the idea that metadata records would use Dublin Core together with other specialized vocabularies to meet particular implementation requirements. During that time, the World Wide Web Consortium's work on a generic data model for metadata, the Resource Description Framework (RDF), was maturing. As part of an extended set of DCMI Metadata Terms, Dublin Core became one of the most popular vocabularies for use with RDF, more recently in the context of the Linked Data movement. Each Dublin Core element is optional and may be repeated. The DCMI has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements. The Dublin Core became ISO 15836 standard in 2006 and is used as a base-level data element set for the description of learning resources in the ISO/IEC 19788-2 Metadata for learning resources (MLR) – Part 2: Dublin Core elements, prepared by the ISO/IEC JTC1 SC36. Full information on element definitions and term relationships can be found in the Dublin Core Metadata Registry.

The DOI's (Digital Object Identifier, www.doi.org/doi_handbook/4_Data_Model.html#4.3) approach to metadata has two aspects: first, the DOI standard mandates a particular mini-

mum set of metadata (the "Kernel" metadata) to describe the referent of a DOI name, supported by an XML Schema; secondly, to promote interoperability and assist RA's in the creation of their own schemas the IDF provides a Data Dictionary or ontology of all terms used in the Kernel, and other terms registered by Registration Agencies, and supports a mapping tool called the Vocabulary Mapping Framework. The "DOI Kernel" is a minimum metadata set with two aims: recognition and interoperability.
Recognition in this context means that the Kernel metadata should be sufficient to show clearly what kind of thing which is the DOI referent (by various classifications), and allow a user to identify with reasonable accuracy the particular thing (by various names, identifiers and relationships). These two are complementary, for it is possible to know that something is e.g. a movie or a DVD without knowing that it is "Casablanca", and vice versa. Recognition is required for the discovery of referents, and also to provide information to a user when a referent is discovered, whether by intent or accident. The user of metadata may be a person or a machine. The structure of the Kernel is often but not always sufficient to provide a unique description of a referent (*disambiguation*), and further specialized metadata elements may be required in some cases. A unique description can in fact always be achieved by adding additional descriptive text to a referentName, but this is not a satisfactory way if the additional text is being used in place of a formal classification, measurement, identifier, time or other structured contextual metadata, as it undermines the second goal of interoperability.
Interoperability in this context means that Kernel metadata from different DOI Registration Agencies may be combined or queried by the same software without requiring semantic mapping or transformation. Interoperability is achieved when data elements or their values are common to diverse metadata schemas. The Kernel provides this directly by mandating a common set of core elements and classifications, but this of course supports only limited interoperability.
The assignment of a DOI name requires that the registrant provide metadata describing the object to which the DOI name is being assigned. At minimum, this metadata shall consist of a DOI Kernel Metadata Declaration (also known as the DOI Kernel). A specification of data elements (with sub-elements, cardinality, etc.), current allowed values and XML expression is maintained by the IDF (the ISO 26324 Registration Authority). Registration Agencies are expected to ensure that, at a minimum, a DOI Kernel Metadata Declaration is made for each DOI name issued. This may be done in two ways: either a Declaration can be made using the DOI Kernel XSD, or (more usually) the elements of the DOI Kernel can be incorporated into a wider metadata schema issued by the Registration Agency. A Registration Agency has the option of not producing DOI kernel metadata unless asked, i.e. it may convert on demand from an internal representation. The minimum set of metadata a registrant should be concerned with is the minimum that will meet its business requirements, not the technical minimum of the Kernel which will always be much smaller. The Kernel schema makes very few data elements mandatory. A minimum set is a necessary but not sufficient requirement in considering the question of what data a registrant may need to communicate to supply chain partners.

The DataCite Metadata Schema (https://schema.datacite.org/) is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. The resource that is being identified can be of any kind, but it is typically a dataset. There are three different levels of obligation for the metadata properties in this schema:

- Mandatory (M) properties must be provided (e.g., Identifier, Creator, Publisher, Title),
- Recommended (R) properties are optional, but strongly recommended for interoperability (e.g., Subject, Date, Description, Geolocalization);
- Optional (O) properties are optional and provide richer description (e.g., Size, Format, Version, Rights).

To enhance the prospects that data metadata will be found, cited, and linked to original research, it is strongly encouraged to submit the Recommended as well as Mandatory set of properties. Together, the Mandatory and Recommended set of properties and their subproperties are especially valuable to information seekers and added-service providers, such as indexers. The Metadata Working Group members strongly urge the inclusion of metadata iden-

tified as Recommended for the purpose of achieving greater exposure for the resource's metadata record, and therefore, the underlying research itself.

Concerning metadata specifically tailored to describe (e.g., give additional data and information about) linguistic digital content, dedicated metadata schemas have been created by the community. The most relevant for the purposes of the SIGN-HUB project are revised in the following. Metadata for language resources and tools exists in a multitude of formats. Often these descriptions contain specialized information for a specific research community (e.g. TEI headers for text, IMDI for multimedia collections).

TEI (the Text Encoding Initiative, www.tei-c.org), is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. The TEI encoding scheme consists of a number of modules, each of which declares particular XML elements and their attributes. Part of an element's declaration includes its assignment to one or more element classes. Another part defines its possible content and attributes with reference to these classes. This indirection gives the TEI system much of its strength and its flexibility. Elements may be combined more or less freely to form a schema appropriate to a particular set of requirements. It is also easy to add new elements which reference existing classes or elements to a schema, as it is to exclude some of the elements provided by any module included in a schema. In principle, a TEI schema may be constructed using any combination of modules. However, certain TEI modules are of particular importance, and should always be included in all but exceptional circumstances:

- The core module contains declarations for elements and attributes which are likely to be needed in almost any kind of document, and is therefore recommended for global use;
- The header module provides declarations for the metadata elements and attributes constituting the TEI header, a component which is required for TEI conformance;
- The text structure module declares basic structural elements needed for the encoding of most book-like objects.

The specification for a TEI schema is itself a TEI document, referred to as an ODD document (from the design goal originally formulated for the system: *One Document Does it all*). Stylesheets for maintaining and processing ODD documents are maintained by the TEI, and these Guidelines are also maintained as such a document. An ODD document can be processed to generate a schema expressed using any of the three schema languages currently in wide use: the XML DTD language, the ISO RELAX NG language, or the W3C Schema language, as well as to generate documentation such as the Guidelines and their associated web site.

OLAC (the Open Language Archives Community, www.language-archives.org) is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources. The OLAC metadata set is based on the Dublin Core metadata set and uses all fifteen elements defined in that standard. To provide greater precision in resource description, OLAC follows the Dublin Core recommendation (being actually an enriched version of it, specifically targeting Sign Languages) for qualifying elements by means of element refinements or encoding schemes. The qualifiers recommended by DC are applicable across a wide range of resources. However, the language resource community has a number of resource description requirements that are not met by these general standards. In order to meet these needs, members of OLAC have developed community-specific qualifiers, and the community at large has adopted some of them as recommended best practice for language resource description.

IMDI (The ISLE Meta Data Initiative, tla.mpi.nl/imdi-metadata, is a metadata standard to describe multi-media and multi-modal language resources. The standard provides interoperability for browsable and searchable corpus structures and resource descriptions. All TLA archiving

tools (Arbil, LAMUS, etc.) are compatible with IMDI: the web-based Browsable Corpus at the Max Planck Institute for Psycholinguistics allows to browse through IMDI corpora and search for language resources. The standard provides interoperability for browsable and searchable corpus structures and resource descriptions with help of specific tools. The project is partly based on existing conventions and standards in the Language Resource community. Another valuable example is the already-cited NGT corpus, which is fully accessible through IMDI.

Finally, CLARIN's CMDI (Component Metadata, www.clarin.eu/content/component-metadata) provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile"). Both are stored and shared with other users in the Component Registry to promote reuse. Each metadata record is then expressed as an XML file, including a link to the profile on which it is based. The CMDI approach combines architectural freedom when modeling the metadata with powerful exploration and search possibilities over a broad range of language resources. As of now, there are two supported versions of CLARIN's component metadata framework: CMDI 1.1 and CMDI 1.2. They are not interchangeable, but CMDI 1.1 metadata can easily be converted into CMDI 1.2. Considering the project's needs, it would be advisable to create CMDI 1.2 metadata, enabling some new features such as enhanced description of relations between resources, and documentation on component, element and attribute level.
Actually, and precisely, there is no such thing as a single CLARIN metadata scheme. Rather, CLARIN proposes a component-based approach: each user can combine several metadata components (sets of metadata elements) into a self-defined scheme that suits his/her particular needs. Nevertheless, the structure is fixed. Each CMDI files exists of 3 parts:

- a (fixed) Header, containing administrative information (e.g., the author of the file, its creation date, and its PID);
- a (fixed) Resources section, containing links to external files (e.g., an annotation file or a sound recording) and/or other CMDI metadata files (to build hierarchies);
- a (flexible) Components section, where the actual components that this profile contains will appear.

Summarizing, for the need of the SIGN-HUB project searching, indexing, and metadata are essential. Searching into digital content can be implemented as Google-like search, which may even work fine for (un)structured text, but a smarter approach is using structure and semantics such as in CLARIN, for metadata, structured data that describes other data. The problems that we see with existing metadata solutions arise since they typically:

- are inflexible: too many, too specific metadata (IMDI) or too few (DC/OLAC) to general metadata;
- use unfamiliar terminology for the reference community;
- allow limited interoperability (both semantic and functional).

CLARIN solution, instead, proposes a joint metadata domain for that allows to build a single metadata catalog and to gather information on resources (statistics). In this, CMDI is not a new metadata schema that should supersede all others, but rather an environment supporting different metadata schema where new interoperable metadata schema can be created to describe new data types (or old data types for new purposes). To do this it is possible to use shareable reusable Metadata Components from a central registry to build a metadata schema with a well-defined syntax and explicit semantics for the metadata elements: a META-SHARE schema that is now available in CMDI.

The CLARIN ERIC network has already been contacted, and a dedicated workshop has been hosted with its technical director, dr. Dieter Van Uytvanck, to discuss the details of the integration of CLARIN resources within the SIGN-HUB project (e.g., pricing model) concerning the usage of CMDI for medata and data description.

d. Providers of Digital Preservation Services

Cloud services can provide easy, automated replication to multiple locations essential for business recovery planning and access to professionally managed digital storage; in addition, the specialists can add access to other dedicated tools, procedures, workflow and service agreements, tailored for digital preservation requirements. Advantages of cloud are, for instance:

- Cloud services can provide easy, automated replication to multiple locations and access to professionally managed digital storage and integrity checking. As a result, data preservation (durability) of digital information can be at least as good (or better) than can be achieved locally;
- Archives can add access to dedicated tools, procedures, workflow and service agreements, tailored for digital preservation requirements via specialist vendors;
- There are potential cost savings from easier procurement and economies of scale, particularly for smaller archives;
- The flexibility of the cloud allows relatively rapid and low-cost testing and piloting of providers.

The most recent and interesting platforms for the purposes of the project are summarized in the following.

Amazon Glacier is an online file storage web service that provides storage for data archiving and backup. It is part of the Amazon Web Services suite of cloud computing services, and is designed for long-term storage of data that is infrequently accessed and for which retrieval latency times of 3 to 5 hours are acceptable. While an unlimited amount of data can be uploaded for storage for free, the pricing structure for downloading data (retrieval) is far more complex. Getting data out of Glacier is a two-step process. The first step is to retrieve the data from Glacier staging area, subject to Glacier retrieval pricing. The next step is to actually download (transfer) the data. Exceeding the hourly allowance at any time results in a peak retrieval overage charge that gets multiplied by the number of hours in a month.

Preservica (preservica.com/edition/cloud-edition) is a fully cloud hosted OAIS compliant digital preservation platform that also includes public access/discovery to allow users to safely share their archive or collection. It is particularly tailored for the preservation and access solution needed by small to mid-sized organizations and consortia, with plans from 1 to 10TB offered.

ArchivesDirect (archivesdirect.org) features a hosted instance of Archivematica with storage via DuraCloud in secure, replicated Amazon S3 and Amazon Glacier storage. It aims at creating standards-based digital preservation content packages that are archived in secure long-term storage.

Arkivum's (arkivum.com) Archive as a Service provides a fully-managed and secure service for long-term data retention with online access and a guarantee of data integrity that's part of its Service Level Agreement and backed by worldwide insurance. It provides the end-to-end managed solution needed to archive data in a secure and accessible manner for the long term.

DuraCloud (www.duracloud.org) is a managed service from DuraSpace. It provides support and tools that automatically copies content onto several different cloud storage providers and ensures that all copies of the content remain synchronized. It offers solutions for on-line backup and sharing, preservation, and archiving.

The services offered by the Preservica provider seems interesting and adequate with respect to the needs of the SIGN-HUB project. Nevertheless, before deciding to go for this solution, it would be needed to discard the option of using integrated resources offering also storage space, capabilities, and services such as CLARIN.

5. Solutions and Services for GIS systems

a. Introduction

The SIGN-HUB digital platform will comprise an interface for Sign Language grammars and another one for the Linguistic ATLAS. These will comprise, among other features, a hybrid system of text, images and video digital content to support navigation of deaf signers, exposing features such as interactive maps, unconstrained search feature combination search, an icon-based system to search for phonemes (e.g. handshapes), and searchable examples.

In this, capabilities for managing geolocalized content and interactive maps will be of paramount importance. This chapter deals with existing technologies and platforms that represent the state of the art for web-based tools to create, delivery, and manage geolocalized content and interactive maps.

In this, it is of relevance to mention again the WALS project, which interacts with Google Maps and is based on GeoJSON, a format for encoding a geographic data structures (RFC 7946).

Services of interested for the SIGN-HUB project to administrate geolocalization of digital content and to provide users with interactive maps are surely Google Maps and Open Street Maps.

Google Maps (maps.google.com) is a web mapping service developed by Google. It offers satellite imagery, street maps, 360° panoramic views of streets (Street View), real-time traffic conditions (Google Traffic), and route planning for traveling by foot, car, bicycle (in beta), or public transportation. Google Maps began as a C++ desktop program designed by Lars and Jens Eilstrup Rasmussen at Where 2 Technologies. In October 2004, the company was acquired by Google, which converted it into a web application. After additional acquisitions of a geospatial data visualization company and a real-time traffic analyzer, Google Maps was launched in February 2005. The service's front end utilizes JavaScript, XML, and Ajax. Google Maps offers an API that allows maps to be embedded on third-party websites and platforms (being for instance used by Airbnb and Linguistic ATLAS as the Algonquian one, www.atlas-ling.ca), and offers a locator for urban businesses and other organizations in numerous countries around the world.

OpenStreetMap (www.openstreetmap.org) is a collaborative project to create a free editable map of the world. Its creation and growth has been motivated by restrictions on use or availability of map information across much of the world, and the advent of inexpensive portable satellite navigation devices. OSM is considered a prominent example of volunteered geographic information. Created by Steve Coast in the UK in 2004, it was inspired by the success of Wikipedia and the predominance of proprietary map data in the UK and elsewhere. Since then, it has grown to over 2 million registered users, who can collect data using manual survey, GPS devices, aerial photography, and other free sources. This crowdsourced data is then made available under the Open Database License. The site is supported by the OpenStreetMap Foundation, a non-profit organization registered in England and Wales.

Rather than the map itself, the data generated by the OpenStreetMap project is considered its primary output. The data is then available for use in both traditional applications, like its usage by Craigslist, OsmAnd, Geocaching, MapQuest Open, JMP statistical software, and Foursquare to replace Google Maps, and more unusual roles like replacing the default data included with GPS receivers. OpenStreetMap data has been favorably compared with proprietary data sources. OpenStreetMap expose the Web Map Framework and Overpass APIs for fetching and saving raw geodata from/to the OpenStreetMap database, which are now used by many websites and platforms (such as Wikipedia itself).

b. Geographic Information Systems

Until a few decades ago, manipulating, synthesizing and representing geographic information was restricted to paper maps and these tasks were limited to manual, non-interactive processes. The exponential improvement in the performance of computer-based technologies and the increasing demand for interactive manipulation and analysis of geographic information have created a need for Geographic Information Systems (GIS). An important characteristic of these GIS systems is that they are more than tools to produce paper maps.

A GIS software is designed to capture, manage, analyze, and display all forms of geographically referenced information. GIS allows to view, understand, question, interpret, and visualize the world in ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts. GIS is becoming essential to understanding what is happening and what will happen in geographic space. GIS software helps users answer questions and solve problems by looking at data in a way that is quickly understood and easily shared, namely on a map.

In the last years GIS has turned into a compelling Web Application that has prompted many people to take advantage of Internet. The Web has revealed the immense value and broad applicability of GIS and introduced flexible architectures for use with modern IT infrastructure. Web GIS provides a new infrastructure for research. With the evolution of web technologies and the expanding use of web GIS, a large amount of remote-sensing data, volunteered geographic information, and analytic web services are published on maps daily. In addition, there is an exponential increase in the number of sensors directly connected to the web, providing rich, real-time datasets. As large enterprise organizations offer geospatial computing in the cloud through their web services and small organizations and the general public add geospatial services to the web, there is a growing wealth of powerful analytic capabilities available to scientists. They can now assemble the resources they need through web programming interfaces without being highly trained specialists in federated computing, requiring the access to disparate databases to create virtual ones. With the web as an infrastructure for e-science, the entry costs to working digitally have been greatly reduced, making it more readily available to the community[19].

Similarly, the SIGN-HUB community, through the implementation of the interface for the Linguistic Atlas, will provide a comparative description of the linguistic structures of several sign languages taking advantage of Web GIS application, or at least of existing services to set-up an interface for geolocalization of digital content, since it will be too expensive to build a custom solution for this interface.

Figure 4 - Web GIS Application Interface

[19] *About creating web GIS applications* Online at http://server.arcgis.com/en/server/latest/create-web-apps/windows/about-creating-web-gis-applications.htm , *2016.*

A Web GIS Application is a type of distributed information system, comprising at least a server and a client, where the server is a GIS server and the client is a web browser, desktop application, or mobile application. In its simplest form, web GIS can be defined as any GIS that uses web technologies to communicate between a server and a client.

There are four key elements in every web GIS application (Figure 5), all of which can be accessed through the web or via a local area network (LAN):

- *GIS Client*: renders geospatial information in a GeoBrowser;
- *Web MapServer:* performs the requested GIS operations and sends responses to the client via HTTP;
- *Web Server*: hosts the web application that provides the software interface to the client;
- *GeoDatabase:* is responsible to store and make persistent geospatial data.

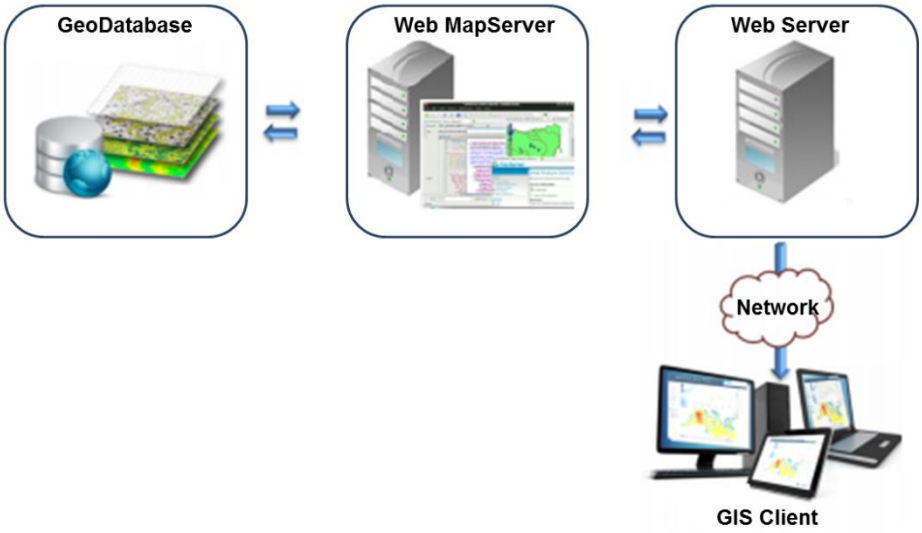

Figure 5 - Web GIS Application Architecture

By utilizing the Internet to access information over the web without regard to how far apart the server and client might be from each other, web GIS introduces the following advantages over traditional desktop GIS:

- Increased accessibility: due to the global nature of web GIS inherited from http, web GIS applications can be presented to the world, and the world can access them from their computers or mobile devices;
- A large number of users: In general, a traditional desktop GIS is used by only one user at a time, while a web GIS can be used by hundreds of users simultaneously;
- Easy to use: Web GIS is intended for a broad audience, including public users who may know nothing about GIS. Web GIS is commonly designed for simplicity, intuition, and convenience, making it typically much easier to use than desktop GIS.

In the following, we describe all the elements that developers must assemble as part of each web GIS application to provide end users with a web GIS application that enables them to get their work done without having to learn a lot about GIS. This is given as a general introduction to the reader to better understand the technical challenges one has to cope when dealing with GIS applications. For the purposes of the SIGN-HUB project, we will deal only with the development of a custom GIS client leveraging on the APIs exposed by the JavaScript library OpenLayers; the web map server and web server will be derived by the external services of OpenStreetMaps; the geodatabase will instead be a SQL database populated with the da-

ta coming from the description of the grammars and the linguistic content as provided by the content creators and the end users answering to the surveys.

OpenLayers (openlayers.org) is an open source (provided under the 2-clause BSD License) JavaScript library for displaying map data in web browsers. It provides a large set of APIs for building rich web-based GIS applications. OpenLayers makes it easy to put a dynamic map in any web page. It can display map tiles, vector data and markers loaded from any source. OpenLayers has been developed to further the use of geographic information of all kinds. It is a free solution,

c. GIS Client

The GIS client tools are responsible to render information on a navigable 2D or 3D map to the users, and to allow them to interact with displayed data. The dynamic rendering and user interaction falling on the client side can be implemented using *Operational layers.*

Operational layers are the small set of layers that users work with directly or derive as the result of an operation (such as a query) in a web GIS application. These layers are often tailored to a particular group of users by a GIS professional. For example, an urban planner uses a Windows smartphone running a GIS application to update the location of manhole covers in a sanitary sewer/storm water system layer. In most GIS applications, users work with operational information (sometimes multiple operational layers) on top of their basemap, which provides the geographic context. At other times, the operational layer is displayed underneath other layers that help provide locational context.

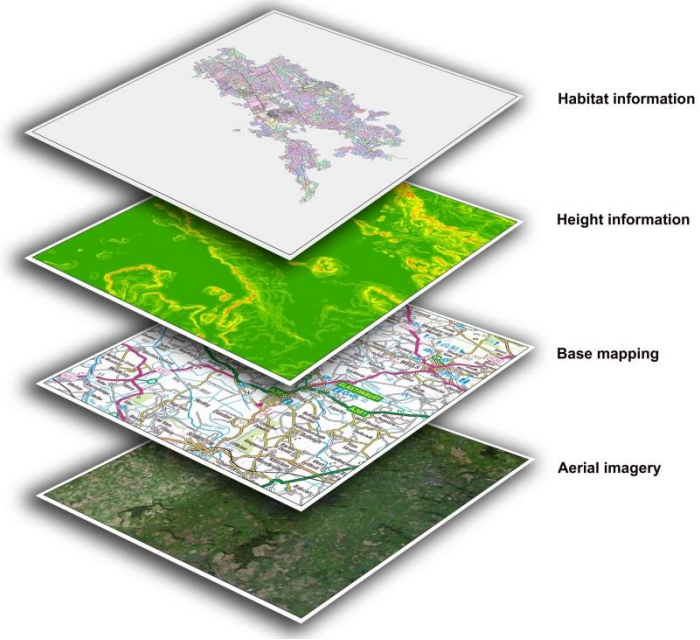

Figure 6 - GIS Layers

For example, when users classify and display ZIP or postal code areas by demographic information, they often overlay these results with transportation lines and place-names to provide locational context. Operational layers are often dynamic; they are retrieved from the GIS database and displayed during runtime, for example, each time users pan, zoom, or refresh a map. It is common that operational layers work within a focused range of map scales and resolutions.

Operational layers may consist of: observations or sensor feeds often displayed as status information in web GIS maps or frequently used as inputs into analytic operations that are computed on the server, that is any information that reflects status or situational awareness, e.g, crime locations, traffic sensor feeds, real-time weather, disease locations, air quality and pollution monitors; query results displayed as map graphics in the web GIS application when applications make a query request to the server and return a set of records as results; result layers that are derived from analytic models when GIS analysis can be performed to derive new information that can be added as new map layers and explored, visualized, interpreted, and compared by end users.

There exist many tools that have been designed in order to provide GIS Client services, such as OpenLayers, Google Maps, Google Earth, ArcGIS Desktop, and Bing Maps.

d. Web map server

The map services are software package or program able to retrieve spatial data and render raster and vector maps image in a GeoBrowser. Since the standard web servers and browsers supports only standard image and data formats like .jpeg, .gif, etc., to represent or publish geo-spatial data in web compatible format there is a need of intermediate software components called as a map server or GIS server.

The aim of map servers is to access existing geospatial information in diverse formats and serve this information to map clients through standard protocols such as WMS protocol. The Web Map Service (WMS) protocol (OGC, 2006) is responsible of dynamically producing maps for georeferenced data from one or more distributed geospatial databases. A WMS request defines the geographic layers and area of interest to be processed, the response is one or more a digital image (JPEG, PNG, GIF) representing a map to be displayed on a web-client. The WMS standard defines three operations: one returns service-level metadata; another returns a map and an optional third operation returns information about particular features shown on a map. Those operations are implemented defining three HTTP requests, respectively: *getCapabilities, getMap* and *getFeatureInfo* requests. All these requests could be used in order to create a basic map where the user can identify a feature, get some basic information from it and perform some basic queries.

There are different types of geospatial information services:

- Map services: receive requests from clients, access vector information (graphic and alphanumerical) and/or raster, and generate images of maps which are served to the clients in the form of raster images or respond to requests to access information relating to the maps served, such as the properties of an element;
- Geometry services: serve geospatial information in vector format, including geometry and alphanumerical attributes, so that the client can process and represent or manage them locally;
- Coverage services: serve different types of geospatial information to the clients, with the particular feature of offering the original information without the need for it to be processed. This is useful for accessing data such as digital models of the terrain (DTM), remote sensing and information archives, etc.
- Catalogue services: offer access to metadata and the search for information about cartography.
- Processing services: offer the processing of geospatial information on the server's side, and the final results are then sent to the clients. (e.g. route calculation, analysis, etc.).

All of these services are usually associated with interoperability standards specified by the OGC (Open Geospatial Consortium).

There are numerous of free and/or open source web map servers available that are remotely hosted or easy to set up, such as GeoServer, MapServer and ArcGIS Server.

e. Web server

A web server is a computer system that processes requests via HTTP, the basic network protocol used to distribute information on the World Wide Web. The term can refer either to the entire system, or specifically to the software that accepts and supervises the HTTP requests. The primary function of a web server is to store, process and deliver web pages to clients.

In the web GIS application architecture, the use of web server is to host the web application that provides the software interface to the client, and its corresponding tools used to visualize, interact with, and work with geographic information. It may be an application that runs in a web browser, or it could be a mobile application that works on a GPS-enabled field device or a smartphone, such as an iPhone.

Users have a number of application choices that can use to build each web GIS. Often, the right choice depends on the set of functions, tools, and map displays required by the users' workflows. Just as often, the choice of application will depend on the end user and his or her experience using computers and the setting in which the work is done.

f. Geodatabase

The geodatabases are responsible to store and make persistent spatial data, offering spatial types inspired to the OGC Geometry. The Geodatabases are software systems that use a standard method of cataloging, retrieving, and running queries on data. The DBMS manages incoming data, organizes it, and provides ways for the data to be modified or extracted by users or other programs. Each GIS application depends on a strong geospatial data management framework that can hold the information used to support the web GIS application. This can be one or more geodatabases, a collection of shapefiles, various tabular databases and spreadsheets, CAD files, design files, imagery, HTML web pages, and so forth. GIS datasets must be compiled in unison, harmonized, and integrated to fit together in a geographic framework.

There are many geodatabases designed to support all levels of GIS implementation, from those that support the simplest geodata models to those that are quite sophisticated (PostGIS/PostgreSQL, OracleSpatial, MySQL, ArcGIsServer).

6. Design Considerations

All the above-mentioned web application technologies have been in-depth investigated in relationship to their possible usage within the SIGN-HUB project. We have analyzed the requirements posed by the online platform to be developed within the project and checked whether each of the described technology is able, or not, to meet them. This section provides and summarizes the outcomes of our analysis, following the same structure enforced throughout the whole document.

a. Management and Distribution of Digital Content

In this, the requirements posed by the SIGN-HUB project relate to management and storage of digital content, with particular regards to:

- persistence of data;
- usability;
- metadata creation and management;
- FAIR-compliancy (even considering naming conventions, better described in deliverable D3.11);
- privacy.

Concerning the persistence of data, we must deal with a large amount of data (rough estimation: 50-80TB of data is to be produced by the SIGN-HUB consortium). We still have to identify the best way to guarantee the highest degree of data persistence possible for both data and metadata, and to identify possible repositories (most probably *buying* existing solutions or at least CMSs-building blocks). Surely, we have already decided that an on-line repository will be delivered within the SIGN-HUB project to ensure long-term data persistence and usability.

Concerning usability, we must guarantee to content creators the information uploading (both for data and metadata) and to end users that data is always accessible; then, we must be able to transmit and show data in easy ways and we must guarantee that the procedure to gain access to data is easy.

Concerning metadata, we must be able to deal with all of the information needed by the content creators for present and future research; we must rely on metadata schema that are extendible for which to design the hierarchy (grouping). We must be able to hide/disable groups of metadata (link to Privacy requirements) and, also, we should deal with batch import from external sources, to not require content creators to have to hand-process the information they have already collected. So, we have to identify possible standards as well as possible repositories.

Concerning privacy, we must give two levels of accessibility for both data and metadata: public and private: we must guarantee that we can match the requirements posed by EU laws and National laws, as well as we must guarantee that each data is reachable only by the users who have been granted access for that data. We therefore should identify tools for authentication, either enabling federates single log-in or custom mechanisms, for protecting most sensible data, and also deal with the logistic and legal issues deriving from open access and informed consent forms. There is a strong conflict between reusability and interoperability from one side and anonymous publishing on the other side: since reusability and interoperability mush be guaranteed, the conflict it must be solved by ensuring that all the digital content to be accessed through the SIGN-HUB platform is licensed to be distributed in Open Access. The privacy requirement for the digital content to be accessed through the SIGN-HUB platform is that exact and explicit information (e.g., metadata information about precise age, or residence address) that may link to subjects with impairments are hidden (e.g., only age range is given).

b. Management and Distribution of Surveys

Concerning the interfaces to be offered by the digital platform of the SIGN-HUB project, namely the four software suites identified within the deliverable D3.1 that are:

- the SignGram Blueprint Suite;
- the Atlas of Sign Language Suite;
- the Test Administration Suite for Sign Languages;
- the Digital Archive of Old Signers' Linguistic and Cultural Heritage Suite.

we see that the survey and assessment tool require functionalities that should be implemented custom within the project.
In fact, none of the software reviewed, being an open-source software or a proprietary alternative, is for instance capable of matching some of the requirements stated in D3.1, such as:

- the Survey Editor should allow Content provider to edit questions containing video (videos that should be stored in the SIGN-HUB digital platform);
- the Test Administration Suite for Sign Languages should allow each content provider to use videos from other surveys, once requested and received authorization from the relative owners;
- the Test Administration Suite for Sign Languages should allow the content provider to administer the questionnaire off-line, record the answers and update the platform once back to Internet.

c. Standards for Archiving and Preservation of Digital Content

Concerning the key feature of long-term storage of data, concerning also the needs for preservation standards, resources identification, and management of metadata, which is of paramount importance for the Digital Archive, the project will surely have to provide a custom uniform digital platform, acting as single point of access toward all of the services exposed, which has to be compliant with the requirements stated in D3.1, such as:

- [styling] the web platform should feature modern and user-friendly interface, showing both the project and the EU logo;
- [accessibility] the web platform should provide content accessible to deaf people and hearing-impaired people, in compliancy with the Web Content Accessibility Guidelines 2.0, level AAA;
- [security - Injection] the web platform should not allow injection flaws (e.g., SQL, and LDAP injection occur when untrusted data is sent to an interpreter as part of a command or query) keeping untrusted data separate from commands and queries.

This platform however could expose functionalities that are actually provided by external services: it will be then needed to evaluate the APIs exposed by existing platforms and repositories (e.g., CLARIN) to implement an architecture where each end user actually interacts only with the unique SIGN-HUB digital platform, e.g., by accessing a document there, retrieving its metadata, seeing a video in streaming, but the digital content is actually hosted somewhere else (e.g., again, in one of the CLARIN centers) in a way that is transparent to the end user.

Concerning identification of resources, the usage of DOI seems to be a valid option. On the other hand, also the exploitation of identifiers specifically tailored for digital linguistic resources has been carefully considered. In this, CLARIN persistent identifiers represent the standard de facto and can be leveraged on for the development for the SIGN-HUB project.

Concerning metadata, the problems that we see with most of the existing metadata solutions derive from them:

- being inflexible: too many, too specific metadata (IMDI) or too few (DC/OLAC) to general metadata;
- using unfamiliar terminology for the reference community;
- allowing limited interoperability (both semantic and functional).

CLARIN solution, instead, proposes a joint metadata domain for that allows to build a single metadata catalog and to gather information on resources (statistics). In this, CMDI is not a new metadata schema that should supersede all others, but rather an environment supporting different metadata schema where new interoperable metadata schema can be created to describe new data types (or old data types for new purposes). To do this it is possible to use shareable reusable Metadata Components from a central registry to build a metadata schema with a well-defined syntax and explicit semantics for the metadata elements: a META-SHARE schema that is now available in CMDI.

It is highly recommendable that, to maximize easy of usage, data interoperability and long-term coherence, only one service provider and repository is identified and used to build the SIGN-HUB dataset (e.g., it is not recommendable to store part of the data in CLARIN, part in EUDAT, part in Zenodo…). In this, the choice to make is whether to go for data depositing or proper storing. In this, some of the the platforms mentioned earlier in this document seem suitable to host the project data, at least in part; in addition, building a custom FAIR-compliant repository ex-novo seems, given the present state of the art, quite an unreasonable choice, which will lead to consistent expenses for the project and would drag resources (e.g., in terms of money and time) out from the development of the interfaces and tools that represent really the main outcomes of the technical Work Package WP3. However, none of them seems to ad-

here totally to the requirements and the challenges posed by the SIGN-HUB project. At least a responsible for all of them has been contacted to explain the needs for the SIGN-HUB project and to negotiate about providing all the services needed. The definitive statements about which platform to use, if any, and which services are to build custom within SIGN-HUB will be made before December the 6th 2017.

d. Solutions and Services for GIS Systems

Concerning the Linguistic Atlas, it will require to handle geolocalized content. In this, it requires functionalities that should be implemented custom within the project, although the services exposed by map provides such as OpenStreetMap and the APIs exposed by the OpenLayers library for displaying map data in web browsers will be surely used to speed up the development process.

7. Conclusion

The Deliverable 3.2 gives an overview of the existing web technologies that are usually adopted to develop a web application. According to the project objectives related to the development of SIGN-HUB software platform, we especially focused on the software and services that can be worth for SIGN-HUB platform development. Therefore, it is provided a deep analysis of most used CMSs, platforms, repositories and tools providing a final comparison between them focusing on pros and cons. Moreover, a short description about web technologies and GIS tools is also provided. Finally, the results of the analysis related to the design considerations are reported.

In conclusion, we can say that the online digital platform required by SIGN-HUB, as for the requirements stated in deliverable D3.1, needs to expose different services (e.g., survey tool, SignGram Blueprint, multimedia content, etc.) to a variegate and large group of users, of which we should assume that most of them are not at ease with technology.

So, it is of paramount importance for any enabling technology of the online platform in the SIGN-HUB project to be of simple usage and to not require excessive tuning (i.e., even the default options and settings should guarantee an optimal user experience).

Finally, the conclusion of the *build or buy* analysis and the related activities as described in this document is that no existing tool meets fully the requirements and desiderata from our users. Then, we will go for using the fundamental technologies (we will rely on HTML, Javascript, and Java, as detailed in deliverable D3.5) to implement the online platform within the project: its outcomes will be novel and provide benefit to the community. Nevertheless, we plan to leverage on existing services (e.g., OpenStreetMap for the GIS application), knowledge and features (e.g., the Survey tool will implement drag and drop features mutated, although simplified, from existing applications as Qualtrics), and above all on existing repositories for ensuring long-term data persistence.

In particular, we have already contacted the responsible people from both the EUDAT and the DARIAH network. Then, many discussions have already been taken with the CLARIN ERIC network; a dedicated workshop has been hosted with its technical director, dr. Dieter Van Uytvanck, to discuss the details of the integration of CLARIN resources within the SIGN-HUB project (e.g., pricing model) concerning the usage of CLARIN repositories and the possibilities from CLARIN centers to host SIGN-HUB data. The definitive statements about which platform to use, if any, and which services are to build custom within SIGN-HUB will be made before December the 6th 2017. The choices and the final decision will be provided on the basis of a cost analysis (best value for money will be guaranteed) and of an in-depth review of the interfaces provided by these platforms in terms of accessibility for both end-users (e.g., human-machine interface) and our digital platform (e.g., APIs to retrieve their content from the digital platform of SIGN-HUB).

The first action to take in the near future for us is then certainly to identify a platform and a repository for archiving the digital content already provided in WP2, especially video and its metadata, and understand how it can be linked to the SIGN-HUB platform. Then, the APIs exposed by the chosen platform will be leveraged to implement the digital architecture needed for the data exchange and communication between the external platform and the SIGN-HUB one. In parallel to this, WP3 has already started an in-depth analysis with WP2 leaders to investigate 1) the status of the digital material already collected, 2) what is left to be collected, and 3) potential ethical issues that emerged after the project started.