eXascale Infolab

SWITCH

SWITCH, Werdstrasse 2, P.O. Box, CH-8021 Zürich
www.switch.ch

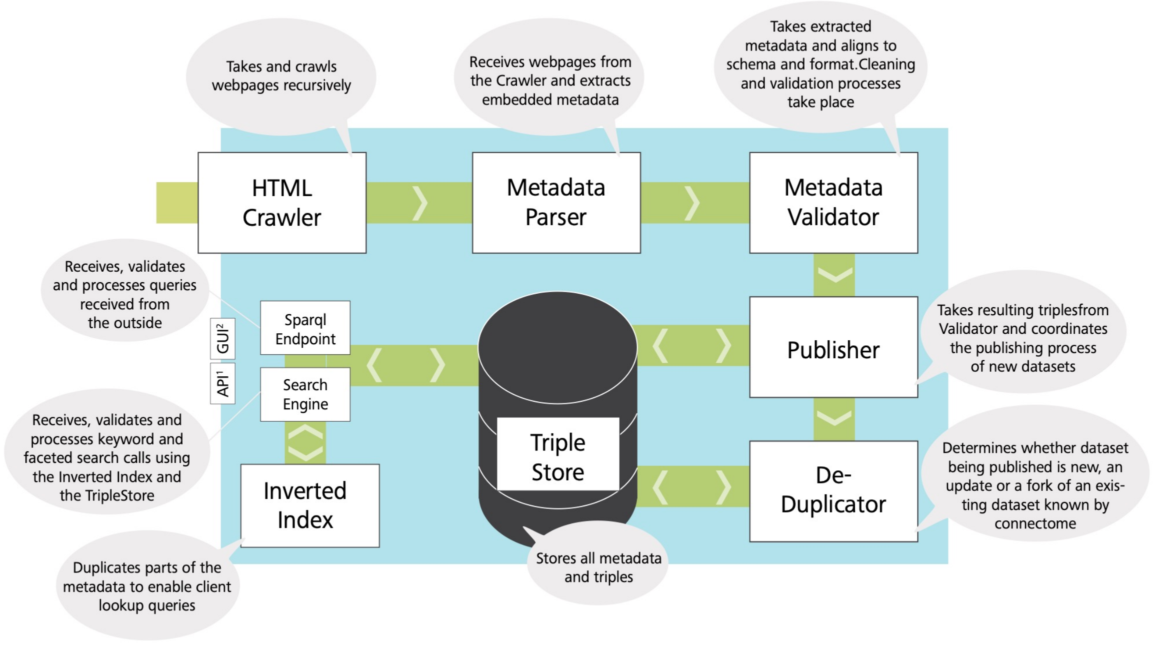
SWITCH Innovation Lab “Research data connectome technologies”

Innovation Lab Partner
eXascale Infolab at Université de Fribourg - on behalf of SWITCH
Philippe Cudré-Mauroux

Executive Summary

Up-to-date, freely accessible and usable research data is the most important resource for ensuring Switzerland's strengths in Research and Innovation and a catalyst for the development of new research findings. Scientists across disciplines generate increasing amounts of data as a byproduct of their daily research activities. Being able to reuse or even combine such scientific data opens the door to many exciting possibilities (improved reproducibility, faster discoveries, more comprehensive studies and models, etc.).

This report proposes a high-level architecture for the research data connectome, a vision initiated by SWITCH, that could make research data findable, accessible, reusable, and more connected. The results presented focus on the current status of best practices from research and industry regarding the publication and interlinking of research data, as well as architectural considerations for a potential connectome. The architecture for a minimal viable product for the connectome is proposed as follows:


Takes and crawls webpages recursively
Receives webpages from the Crawler and extracts embedded metadata
Takes extracted metadata and aligns to schema and format.Cleaning and validation processes take place
HTML Crawler
Metadata Parser
Metadata Validator
Receives, validates and processes queries received from the outside
GUI²
API¹
Sparql Endpoint
Search Engine
Publisher
Takes resulting triplesfrom Validator and coordinates the publishing process of new datasets
Receives, validates and processes keyword and faceted search calls using the Inverted Index and the TripleStore
Triple Store
Inverted Index
De-Duplicator
Determines whether dataset being published is new, an update or a fork of an exis-ting dataset known by connectome
Stores all metadata and triples
Duplicates parts of the metadata to enable client lookup queries

[1] GUI layer provides researchers with the interface to access the Connectome
[2] API layer provides machines and systems with interface to access the Connectome

A high-level architecture of the connectome, Prof. Philippe Cudré-Mauroux, eXascale InfoLab, on behalf of SWITCH Innovation Lab "Technologies for a research data connectome"

Design Considerations on SWITCH's Connectome Vision

Prof. Dr. Philippe Cudré-Mauroux
eXascale Infolab
U. of Fribourg—Switzerland

A SWITCH Innovation Lab
managed by Dr. Sebastian Sigloch

February 17, 2020

Abstract

Scientists across disciplines generate increasing amounts of data as a byproduct of their daily research activities. Being able to reuse—or even combine—such scientific data opens the door to many exciting possibilities (improved reproducibility, faster discoveries, more comprehensive studies and models, etc.) In this report, I summarize my work on SWITCH's vision for a Research Data *Connectome*, a software system that could make Swiss research data findable, accessible, reusable, and more connected. The results presented below focus on the current status of best practices from research and industry regarding the publication and interlinking of research data, as well as architectural considerations for a potential Connectome platform.

1 Importance of a Research Data Connectome

Researchers increasingly produce large quantities of data as a byproduct of their daily activities. Still, it is very difficult for them to share that data, or to find and reuse data from other scientists (esp. across multiple domains). The data sharing landscape is highly fragmented, with thousands of research data repositories [1], and the way to serialize, describe, and offer the data can vary widely from one repository to the next. This creates a complex ecosystem where scientists are unsure where to publish their data on one hand, and where they cannot discover third-party data they are searching for [2] on the other hand.

Tackling this problem is especially crucial today, as Big Data Integration techniques have matured [3] and as models have become more data-hungry (e.g., the rise of deep learning techniques asks for very large and curated training, test, and validation sets). There is a clear need for offering better tools to publish, search and reuse scientific data; this need recently crystallized with the rapid rise of the FAIR principles for data publishing (see below Section 2.2) emphasizing machine-actionability (i.e., software platforms should be able to locate and parse data automatically) and the creation of new tools making scientific data Findable, Accessible, Interoperable, and Reusable.

SWITCH's vision for a research data *Connectome*[1] is an answer to that issue. It envisions a software platform to make Swiss research data FAIR-er, where data from various disciplines—stored in different locations—would become easily findable, accessible, reusable and connected. The present report summarizes the state of the art in this context, and proposes design principles to realize a first implementation of this vision.

2 Identification of Best Practices

Before recommending a potential solution, I give below a brief overview of current best practices related to the Connectome. I start by summarizing the main insights from a series of interviews I conducted, before describing the main standards and platforms relating to our efforts.

2.1 Interviews with Stakeholders

I interviewed researchers from several disciplines in order to understand the best practices in terms of publishing, and potentially interlinking, research data. In particular, I studied the types of data covered by best practices, how one could increase the findability, reusability and sharing of research data, and what incentives we may derive from researchers in that context. Short transcripts of some of those interviews are available in the Appendix. The summary of those interviews are as follows:

- at this point, there is no agreed-upon standard for publishing research data in Switzerland (or abroad, as far as I know); data publication varies depending on the discipline and the individuals, and is mostly ad hoc;

- very few researchers link their data to further data at this point, mostly because of the complexity of the current (e.g., Linked Data) platforms, or simply because the scientists have not been exposed to the practice of linking data;

- when scientists publish data, they usually publish it in raw format with limited metadata, on their website of using well-known platforms like Zenodo (see below Section 2.3);

- scientists are eager to publish their data, to foster further research as well as to gain recognition. However, the lack of time and the complexity of the current tools are often cited as hurdles in that context.

[1] `https://www.switch.ch/stories/research-data-connetcome/`

- legal concerns (e.g., on the rights to distribute, use, and build upon the share data) are an issue for many researchers when opening up their data.

2.2 Relevant Standards

A number of guidelines and standards have been proposed in the past to foster the reuse of scientific artifacts and to encode scientific metadata. I briefly describe the most relevant initiatives in this context below.

2.2.1 FAIR

FAIR [4] is a recently published set of guidelines and principles to make scientific data Findable, Accessible, Interoperable, and Reusable. The authors note in their original paper that there is an urgent need to improve the infrastructures supporting the reuse of scholarly data, and suggest a set of principles towards that goal as follows:

To be **Findable**:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be **Accessible**:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1. the protocol is open, free, and universally implementable

A1.2. the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be **Interoperable**:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be **Reusable**:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards.

The FAIR principles are rapidly gaining momentum, and more and more initiatives tend to promote them. However, they are still evolving, not always measurable and, according to one of their main authors, not readily "implementable" at the time of writing.

Relationship to Connectome: The Connectome aims to provide a FAIR solution to Swiss researchers. It has the potential to support all principles (A1. to R1.3.) expressed above.

2.2.2 RDF/OWL/SPARQL

The Resource Description Framework (RDF)[2] is a family of specifications for metadata models. It is maintained by the W3C, the main international standards organization for the World Wide Web. RDF encodes metadata as *triples*: *(subject, predicate, object)*, where the subject is an entity (URI) about which the metadata is about, the predicate is a property (i.e., relation, also a URI), and the object is a URI or literal for the value of the metadata, for instance:

```
(http://dbpedia.org/resource/J._R._R._Tolkien,
 http://dbpedia.org/ontology/author,
 http://dbpedia.org/resource/The_Hobbit)
```

to express the fact that J.R.R. Tolkien is the author of The Hobbit. Sets of RDF triples form graphs of data (a.k.a. *Knowledge Graphs*) since they often share the same URIs. RDF can be serialized in various ways, for instance using XML or JSON. Vocabularies can be defined for RDF using RDF Schemas[3] or the Web Ontology Language (OWL)[4]. The latter can be used to express ontologies supporting logical inference (e.g., using decidable fragments of first-order logic). RDF data can be queried by using SPARQL[5], a declarative query language similar to SQL.

Relationship to Connectome: RDF is the most prominent standard to encode online metadata and should probably be adopted by the Connectome.

2.2.3 DublinCore

DublinCore[6] is a widely-used schema to encode metadata relating to digital or physical items (e.g., videos, books or artworks). The original standard included 15 elements (*Contributor; Coverage; Creator; Date; Description; Format; Identifier; Language; Publisher; Relation; Rights; Source; Subject; Title;*

[2] https://www.w3.org/RDF/

[3] https://www.w3.org/TR/rdf-schema/

[4] https://www.w3.org/OWL/

[5] https://www.w3.org/TR/rdf-sparql-query/

[6] https://www.dublincore.org/

Type), which were later extended with *qualified* elements. DublinCore is today often expressed using RDF.

Relationship to Connectome: DublinCore is the most popular schema for describing digital assets and could be used as part of the Connectome.

2.2.4 W3C's DCAT

DCAT[7] is a schema (an *RDF vocabulary*) to encode metadata pertaining to datasets. It is published and maintained by the W3C, the main standardization body of the Web.

DCAT is based on 6 main classes, as follows:

dcat:Catalog representing a catalog, i.e., collections of metadata about datasets or data services;

dcat:Resource representing a dataset, a data service or any other resource that may be described by a metadata record in a catalog; this is an abstract superclass mostly;

dcat:Dataset representing a dataset, i.e., a collection of data, published or curated by a single agent;

dcat:Distribution representing the distribution mechanism of a dataset such as a downloadable file;

dcat:DataService representing the API calls that can be used to access one or more datasets;

dcat:CatalogRecord representing an item in the catalog, primarily encoding registration information (e.g., publication date of a dataset).

The overall schema defines more than 100 RDF classes and properties to encode all records. It can be used in conjunction with other RDF schemas such as VOID [8] to express statistics and additional metadata for RDF datasets.

Relationship to Connectome: DCAT could be used as a standard to express the datasets metadata, although schema.org might also used for that role as it is more widely deployed.

2.2.5 schema.org

Schema.org[9] is a standardization effort founded by major search engines companies (Google, Yahoo, Microsoft, and Yandex to create, maintain and foster the use of schemas on the Internet. It is mostly used to add rich metadata to webpages that can then be parsed by search engines (over 10 million sites use

[7] `https://www.w3.org/TR/vocab-dcat-2/`

[8] `https://www.w3.org/TR/void/`

[9] `http://schema.org/`

such rich metadata according to schema.org's webpage). Its data model is derived from RDF, and the standard supports various serialization formats (e.g., Microdata, RDFa, or JSON-LD).

Schema.org provides a range set of schemas to describe events, organizations, places, products or reviews. Of particular interest to the Connectome, schema.org defines a *Dataset* class[10] to describe datasets, as well as a *DataCatalog* class, to describe collections of datasets. The definition of the *Dataset* class contains around 100 properties to describe datasets, ranging from *title*, *abstract*, or *copyrightHolder* to *encodingFormat*, *licence* or *temporalCoverage*. The schemas link to other vocabularies such as Dublin Core. This class is currently being extended to cover the idiosyncrasies of specific research domains (e.g., the newly-defined DATS [5] schema suggested extensions for medical data as part of its JSON-LD serialization).

Relationship to Connectome: schema.org is already widely used today to describe datasets (see below Google Dataset Search) and would probably be a good starting point to describe the data interlinked by the Connectome. Schema.org's governance is also noteworthy in the context of the Connectome. It operates through two governing bodies: a larger *community group* discussing potential improvements, and a small *steering group* approving release candidates. Most discussions, decisions and releases are handled through online forums and github's issue tracker[11].

2.2.6 Open Archives Initiative Protocol for Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[12] is a protocol to share metadata across repositories. It basically works in a client-server fashion; the server asks for metadata over HTTP, and the client answer using DublinCore metadata in XML, with additional elements optionally attached.

Relationship to Connectome: OAI-PMH could be used by the Connectome as a protocol to share metadata across repositories. Other mechanisms could be used, however, such as crawling JSON-LD metadata or using a SPARQL endpoint.

2.3 Existing Solutions

There already exist many solutions today for sharing research data. Few of those solutions support the interlinking of independently published data, however. I briefly describe below the most important platforms in our context.

[10] http://schema.org/Dataset

[11] https://github.com/schemaorg/schemaorg/issues

[12] https://www.openarchives.org/pmh/

2.3.1 linkub.ch

Linkub.ch[13] provides "an environment to support academic and administrative studies based on linked data respecting personal data protection and scientific principles". The recent project is the result of a collaboration between FORS[14], TREE[15], and the "on the move" NCCR[16]. It proposes to develop a national strategy to provide an environment adapted to the production and use of linked data. The service will be available in 2020 and will basically offer three functionalities:

Metadata: linkhub.ch will provide a metadata service with information on datasets and how they could be accessed and linked data;

Linking: linkhub.ch will provide data linking services that support combining information from different data sources;

Data Access: linkhub.ch will provide storage facilities with information and access to linked data in a secure environment.

Relationship to Connectome: The project might offer services that are similar to those of the Connectome, thought its exact features (e.g., search capabilities) are unclear at the time of writing.

2.3.2 GeRDI

The Generic Research Data Infrastructure (GeRDI)[17] is a research project started in 2016, running for 3 years and financed by the DFG[18] with about 3M EUR. The goal of the project is to "provide generic, sustainable and open software connecting research data repositories to enable multidisciplinary and FAIR research data management".

In the first phase of the project, three data centers supporting the management of research data were linked up with each other such that "research data can be used across disciplinary boundaries, enabling new opportunities for multi-disciplinary research". In a second phase, the developed solution will be rolled-out in Germany and − if appropriate funding is available − serve as a model for a future German Research Data Infrastructure. In particular, GeRDI will be able to support universities and research institutes in providing research data, in linking up their existing data stores and in establishing new research data stores.

Relationship to Connectome: The project might offer services that are similar to those of the Connectome, thought the information and software artifacts available on the project's website at the time of writing are quite limited.

[13] https://linkhub.ch/

[14] https://forscenter.ch/

[15] https://www.tree.unibe.ch/index_eng.html

[16] https://nccr-onthemove.ch/

[17] https://www.gerdi-project.eu/

[18] https://www.dfg.de

2.3.3 figshare

Figshare[19] is an online repository where researchers can make their research output available in a citable and discoverable manner. It allows users to upload any type of data (up to 5GB per item), creates identifiers (DOIs) for the data and supports limited metadata (categories, keywords, licence). The platforms offers a range a simple features such as keyword search, data citation export, or sharing through social media platforms.

Relationship to Connectome: The project might offer services that are similar to those of the Connectome, although the functionalities it offers are quite limited at this point (no rich metadata, simple search, no interlinking, etc.)

2.3.4 CKAN

The Comprehensive Knowledge Archive Network (CKAN[20]) is a web-based, open-source management system for the storage and distribution of open data. It is a powerful data catalogue system that is mainly used by public institutions and governments seeking to share their data with the general public (e.g., to power `data.gov` or `data.gov.uk`). CKAN is also used as the back-end for DataHub[21], one of the most important portals of open data.

Each dataset hosted on a CKAN installation is given its own page for the listing of data resources and a rich collection of metadata that can be searched. The supported metadata are as follows (according to the website):

Title allows intuitive labelling of the dataset for search, sharing and linking;

Unique identifier dataset has a unique URL which is customizable by the publisher;

Groups display of which groups the dataset belongs to if applicable. Groups (such as science data) allow easier data linking, finding and sharing amongst interested publishers and users;

Description additional information describing or analyzing the data. This can either be static or an editable wiki which anyone can contribute to instantly or via admin moderation;

Data preview preview .csv data quickly and easily in browser to see if this is the dataset you want;

Revision history CKAN allows you to display a revision history for datasets which are freely editable by users;

[19] `https://figshare.com/`

[20] `https://ckan.org/`

[21] `https://datahub.io/`

Extra fields these hold any additional information, such as location data (see geospatial feature) or types relevant to the publisher or dataset. How and where extra fields display is customizable;

Licence instant view of whether the data is available under an open licence or not. This makes it clear to users whether they have the rights to use, change and re-distribute the data;

Tags see what labels the dataset in question belongs to. Tags also allow for browsing between similarly tagged datasets in addition to enabling better discoverability through tag search and faceting by tags;

Multiple formats see the different formats the data has been made available in quickly in a table, with any further information relating to specific files provided inline;

API key allows access every metadata field of the dataset and ability to change the data if you have the relevant permissions via API.

CKAN supports a federated mode where several installations can share data between each other; It also supports the Data Catalog Vocabulary (DCAT, see above) such that other, non-CKAN catalogues can also be *harvested* (i.e., data can be pulled regularly into CKAN from existing repositories).

CKAN is open-source and uses PostgreSQL for storing data and SOLR to power keyword search.

Relationship to Connectome: The project is open-source and the connectome might borrow some of its back-end design and functionalities.

2.3.5 Dataverse

Dataverse[22] is similar to CKAN in the sense that it is an open-source web application to share, preserve, explore, and analyze research data. The project originates from Harvard and also features metadata, used to create data citations such as the following:

```
@data{DVN/QUFRJ5_2019,
author = {MacLeod, W. Bentley and Currie, Janet M},
publisher = {Harvard Dataverse},
title = "{Code and Data for: Understanding Doctor Decision Making}",
UNF = {UNF:6:Uvkor5YExGPEQmMPOgBv4w==},
year = {2019},
version = {V2},
doi = {10.7910/DVN/QUFRJ5},
url = {https://doi.org/10.7910/DVN/QUFRJ5}
}
```

[22] https://dataverse.org/

Such citations are standardized in a *Data Citation Standard*[23].

Dataverse organizes data in virtual archives called Dataverses. Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (potentially including documentation and code accompanying the data)

Relationship to Connectome: The project is open-source and the connectome might borrow some of its functionalities, as well as its citation standard.

2.3.6 DSpace

DSpace[24] is probably one of the oldest pieces of software designed to create open-access repositories for scholarly and digital content. It is an open-source project created in 2002 by MIT and HP Labs. It supports the storage of arbitrary data, simple metadata (stored in a relational database according to subset of the Dublin Core schema) displayed as HTML, as well as faceted search and browsing powered by Solr.

Relationship to Connectome: The connectome could reuse part of its frontend or backend, though the functionalities of DSpace are limited (limited API, limited metadata, no linking).

2.3.7 EPrints

EPrints[25] is an open-source software for building open access repositories. It is developed by the University of Southampton. It supports the storage and indexing and listing of arbitrary data, described through simple metadata. Data citations are supported also.

Relationship to Connectome: Limited; The connectome could reuse part of its frontend or backend, though the functionalities of EPrints are limited (limited API, limited metadata, no linking).

2.3.8 Invenio

Invenio[26] is an open-source project started at CERN to build large-scale institutional repositories of digital assets. It was originally built to support MARC21[27] as the only metadata format, but moved later to a JSON-based system to support various metadata, e.g., DublinCore through the Open Archives Initiative (OAI) metadata harvesting protocol (see above). Invenio is JSON-native, provides RESTful APIs and has been designed from the gound up to manage millions of records and petabytes of files.

Relationship to Connectome: The Connectome could reuse part of Invenio is back-end infrastructure.

[23] `https://dataverse.org/best-practices/data-citation`

[24] `https://duraspace.org/dspace/`

[25] `https://www.eprints.org/`

[26] `https://inveniosoftware.org/`

[27] `https://www.loc.gov/marc/bibliographic/`

2.3.9 Zenodo

Zenodo[28] is a free and open online repository enabling scientists to share and preserve their research outputs (up to 50 GB per record). It was created by OpenAIRE (a Horizon 2020 consortium supported by the EU) and the CERN. Zenodo is a "small layer" on top of Invenio and stores data at CERN. Zenodo's roadmap is shared openly on GitHub[29], and contributions from anyone are welcome. All its metadata are openly available under a CC0 licence, and all its contents are openly accessible through open APIs. It uses PostgreSQL as data back-end, Elasticsearch 2.x for search, Redis for in-memory storage and RabbitMQ for messaging.

Relationship to Connectome: Zenodo is a proven solution that is readily-available and could be used as part of the Connectome's pilot. The fact that the service is hosted by CERN—a European research organization—could (potentially?) be an issue to some Swiss data providers or institutions, however.

2.3.10 RENKU

RENKU[30] is an open-source platform built by the Swiss Data Science Center (SDSC)[31] enabling multidisciplinary data science collaboration. RENKU supports versioning of data and code; it tracks which results were produced by whom and when. RENKU encodes all metadata and provenance information in a knowledge graph that can then be queried, e.g., to find who is using a given dataset and how, or to reproduce a scientific workflow end-to-end.

RENKU is based on GitLab for versioning and features several hooks to seamlessly integrate with data science tools (e.g., for interactive notebooks). It uses schema.org for metadata and PROV-O[32] for expressing provenance. It supports SPARQL 1.1 for querying and authentication via ORCID. It also supports a federated mode where metadata can be harvested across instances.

Relationship to Connectome: Renku is one of the only readily-available platforms supporting provenance for scientific data. It could be used by the Connectome to support Fair R1.2. (metadata are associated with detailed provenance), and potentially as a platform for enabling scientific collaboration beyond the sharing of metadata.

2.3.11 Blue Brain Nexus

Blue Brain Nexus is an open-source solution to support Open Science. It is a data and knowledge management platform designed to enable the ingestion, integration and search of any kind of data. Nexus allows to search, discover, reuse and derive high-quality data generated within and outside the platform. At its core, it encodes all metadata as an RDF knowledge graph, making it

[28] https://zenodo.org/

[29] https://github.com/zenodo/zenodo

[30] https://datascience.ch/renku/

[31] https://datascience.ch/

[32] https://www.w3.org/TR/prov-o/

easy to leverage various schemas (e.g., schema.org, bioschema.org, W3C-PROV) and specialized ontologies. It also supports to constrain/validate the metadata through W3C's Shapes Constraint Language (SHACL)[33].

Nexus is using Cassandra as primary data store, ElasticSearch as inverted index (supporting search functionalities) and blazegraph as triple-store (supporting SPARQL queries). It offer both a web interface (*Nexus Web*), allowing users to interact with the data and perform administration tasks, as well as an API to build custom applications.

Relationship to Connectome: Nexus is a readily-available solution for the ingestion and validation of arbitrary metadata stored as an RDF knowledge graph and as such could be used as back-end for the pilot of the Connectome.

2.3.12 Leibniz Data Manager

The Leibniz Data Manager[34] was developed to "support the aspect of better re-usability of research data". The current prototype supports the management and access to heterogeneous research data publications and assists researchers in the selection of relevant datasets for their respective disciplines.

The prototype currently offers the following features for research data:

- Supports data collections and publications with different formats
- Different views on the same data set (2D and 3D support)
- Visualization of Auto CAD files
- Jupyter Notes for demonstrating live code
- RDF Description of data collections

The prototype is based CKAN (see above). It also uses JupyterHub, PostgreSQL and SOLR. A number of extensions of the prototype are currently in the works.

Relationship to Connectome: The Leibniz Data Manager is close to the vision of the Connectome. The connectome could reuse part of the manager to power its services, though the tool is still under active development.

2.3.13 Google Dataset Search

Google Dataset Search[35] is a dataset-discovery tool provided by Google. It offers a simple-to-use, keyword-based GUI[36] to find datasets published on the Web. The tool relies on Google's main crawler for finding the datasets, by parsing HTML pages and indexing metadata describing datasets and embedded in the webpages.

[33] https://www.w3.org/TR/shacl/

[34] https://labs.tib.eu/info/en/project/leibniz-data-manager/

[35] https://dl.acm.org/citation.cfm?doid=3308558.3313685

[36] https://toolbox.google.com/datasetsearch

The tool supports several schemas for describing the datasets: the *Dataset* and *DataCatalog* from schema.org (which represent 98% of the cases), and the *Dataset* schema from DCAT[37] (2% of the cases). The metadata can either be serialized as RDFa, microdata, or JSON-LD. It collects all triples from those schemas, converts the values to a homogeneous representation (using custom wrappers to align for instance all dates), links prominent entities (e.g., organisations and locations) to their Google Knowledge Graph representation, and stores all metadata from a given page as a separate graph (no interlinking between the various graphs) using a schema similar to schema.org's *Dataset*.

Search is then powered by indexing the resulting metadata and ranking results following Google's algorithm for webpages (mostly, with added signals accounting for metadata quality). The tool identifies duplicates by hashing some values from the metadata. However, the results returned by the current version of the tool were considered very poor by some users (see interviews, where users complain about the ranking of the results and the very bad quality of the data returned).

Several enhancements of this tool are in the works, including: i) faceted search (by "date, location, size of a dataset, licensing terms, and other attributes") ii) indexing raw data itself in addition to the metadata and iii) improving the quality of the metadata by learning from existing metadata and linking to further resources such as academic publications and iv) inferring relationships between datasets (i.e., a simple form of provenance).

Relationship to Connectome: Google Dataset Search is close to the Connectome's vision in the sense that it allows to search and reuse scientific datasets. It is however a closed-source solution that cannot be extended and that misses a number of important features (e.g., SDK, provenance, faceted or structured search, etc.) in our context.

2.4 Learnings

A number of findings can be extracted from the points above. First, integrating or linking datasets is too complex to be handled by the scientists (both because there are too many pieces of data out there and because it is challenging technically). Second, there exist a number of standards that could (and should) be directly leveraged by the Connectome. Third, there exist a number of ready-to-use solutions to publish datasets that the Connectome could interface with. However, few of those solutions propose to connect datasets at this point, which is the main goal of the Connectome.

3 Design Considerations for a Data Connectome

Given the vision of a Swiss data Connectome, the best practices as well as the current solutions described above, many important decisions have to be made

[37] `https://www.w3.org/TR/vocab-dcat-2/`

prior to designing a technical architecture for the Connectome. I summarize the most important decisions below.

3.1 Connecting New VS Old Datasets

The Connectome could be designed to index and interlink *new* datasets only, which would be prepared according to a set of guidelines to be correctly indexed and processed. Alternatively, it could also handle older datasets or datasets not meant to be interlinked, raising additional challenges in terms of creating new links or metadata automatically.

3.2 Registering VS Crawling Sources

The Connectome could discover new data sources in two main ways: by letting scientists register new URIs where they publish data, or by crawling websites in order to automatically discover new sources.

3.3 Handling Metadata VS Raw Data

The Connectome could store (resp. index or interconnect) the metadata (the *schema* in database terms) used to describe the dataset only. Alternatively, it could also handle the raw data (the *instances* in database jargon). Storing, indexing and/or interlinking the raw data poses acute challenges, however, as the format as well as the contents of the raw data can be arbitrary (e.g., data could be shared as Excel files, using specific imaging formats, or in binary format).

3.4 Human VS machine Interface

The Connectome could have humans and/or machines (i.e., other pieces of software) as end-users. The interface that it should expose is dependent on this choice (GUI and human-readable interfaces VS APIs and machine-processable formats).

3.5 Centralized VS Decentralized

The connectome could be designed as a centralized solution, where metadata from various sources are replicated and consolidated into a single query point, or as a decentralized system, where each source installs some software (e.g., a SPARQL end-point) and where queries get answered in a collective, federated way involving several sources.

3.6 Query modalities

The Connectome could support various query modalities. Keyword (i.e., unstructured) query is the only mode supported by the Google Dataset Search, for instance. Many users expect different modalities, however, such as faceted search (search by title, year, or discipline), query by example, structured queries (using a query language such as SPARQL, see Section 2.2), or ad-hoc user interfaces.

3.7 Provenance

Correctly indexing the datasets based on their topic and discipline is not always sufficient. As it is very easy to republish data, one can expect the rapid proliferations of duplicated, cleaned, reprocessed or forked datasets on the Connectome. Capturing the provenance (see FAIR R1.2. in Section 2.2) of each dataset is hence essential, especially in our context where data citation [6] becomes an increasingly important topic for assessing the impact of a researcher.

4 Recommended Architecture

Given all the options above, many different versions of the Connectome could be built. I briefly comment on the technical, legal and governance issues associated to each design decision listed above, before suggesting a minimal viable product and possible extensions for the Connectome.

4.1 Discussion on Design Considerations

4.1.1 Connecting New VS Old Datasets

Connecting new datasets only would require scientists to publish specific metadata according to a given schema. It would also probably require the definition (or extension) of a schema (or sets of schemas) to describe the Connectome's datasets. Connecting older datasets would require the automatic conversion of arbitrary metadata into the Connectome's format, and/or the automatic creation of specific metadata based on arbitrary raw data, which is not feasible technically today. Hence, I recommend to handle new datasets that are properly described only, though some additional schema or data could be handled on case-by-case by developing data wrappers (e.g., to recognize prominent formats already used in specific fields).

4.1.2 Registering VS Crawling Sources

Even though web crawling is a well-studied problem, with many open-source packages implementing it (e.g., Apache Nutch[38]), crawling large portions of the web requires important resources and is a demanding exercise. In addition, crawling specific portions of the web is rather delicate, as identifying specific sources is technically difficult (e.g., crawling Swiss websites only is challenging, as there is no register listing all Swiss websites, and as IP addresses or metadata alone are insufficient to correctly identify such sites). On the other hand, asking scientists to register each new dataset is probably to cumbersome. Hence, I

[38] http://nutch.apache.org/

recommend a hybrid approach, where scientists register a URI (e.g., `https://www3.unifr.ch/inf/`) once using a web page/form, and where the Connectome regularly crawls all pages in the registered subdomains to discover new datasets (i.e., to discover new pages with embedded metadata describing a new dataset). How to give proper incentives to Swiss scientists to register their domains is an open question, but could be potentially promoted or handled by funding agencies.

4.1.3 Handling Metadata VS Raw Data

Indexing and properly interrelating metadata is technically challenging but feasible as Google Dataset Search illustrates. Indexing and interlinking raw data would be desirable from an application perspective, as the metadata typically cannot describe all instance data properly. However, correctly indexing and interlinking arbitrary, raw data is infeasible today (as machines, or even humans, cannot understand all possible data or formats). Also, it would add a high burden in terms of computational power and storage, as parsing large datasets is expensive. Hence, I suggest to handle metadata only, and to potentially index and interrelate raw data in a case-by-case scenario (e.g., for textual datasets). It would be however highly desirable to have a data quality check (either manual or automatic) on the data itself before indexing any dataset (as data quality issues are very prominent on current platforms).

4.1.4 Human VS machine Interface

Swiss scientists are expected to use the Connectome to search for new data sources, so the human-readable interface is a must-have. On the other hand, several partners have already expressed the desire to connect to the Connectome programmatically (e.g., to publish data automatically or to create further applications). Hence, I suggest to support both modes and to support both human-readable and machine-processable interfaces.

4.1.5 Centralized VS Decentralized

Decentralizing the Connectome is technically-speaking complex; even if SPARQL federation could be leveraged, this would put additional burden on the scientists / sources (i.e., to properly setup, and maintain a SPARQL end-point), limit the query modalities of the Connectome (see below), and severely degrade the latency of the system. Hence, I recommend to be conservative and to adopt a centralized approach, where the Connectome regularly harvests and centrally consolidates and serves metadata from various sources.

4.1.6 Query modalities

Both unstructured (e.g., keyword) and structured (e.g., SPARQL) queries are must-haves, given that we want to support both humans and machines (see above). In addition, users have expressed the need for faceted search on Google

Dataset Search. In the context of the Connectome, Faceted search is in my opinion essential, as scientists are organized in specific sub-communities and are used to taxonomies to structure their work. Hence I suggest to support all three modalities for the Connectome: structured, unstructured, and faceted. Additional modalities, such as query by example ("find me datasets resembling this dataset") could potentially be supported in a second step.

4.1.7 Provenance

From a technical perspective, provenance exists in many different flavors (e.g., provenance polynomials in database applications, or workflow provenance in e-Science infrastructures). Generally-speaking, storing, tracking and processing provenance is a very complex endeavor [7], and often adds considerable overhead to the design and deployment of data-centric systems. Tracking provenance end-to-end is outside of the jurisdiction of the Connectome (which only see the final product, i.e., the published dataset). However, the Connectome should be able to expose provenance information capture from other tools (e.g., RENKU, see Section 2.3) and published as part of the data (using a standard like W3C's PROV-O). In addition, the Connectome should be able to offer simple provenance capabilities like de-duplication of datasets, similar to what Google Dataset Search offers (see above).

4.2 Minimal Viable Product

Given the design considerations described above as well as the state-of-the-art presented in Section 2, I suggest below a Minimal Viable Product (MVP) for the Connectome, as well as a number of extensions that could be built in subsequent phases.

The MVP consists in a pragmatic solution to realize the vision of the Connectome in a Swiss context, focusing on the following features:

1. Connecting new datasets only;
2. Crawling metadata from specific subdomains;
3. Indexing metadata only;
4. Interfacing to both humans and machines;
5. Offering a centralized service;
6. Offering rich query modalities (unstructured, structured, faceted);
7. Capturing simple provenance in the form of identifying duplicates.

A high-level system architecture of this MVP is depicted in Figure 1.

The MVP is composed of the following main components:

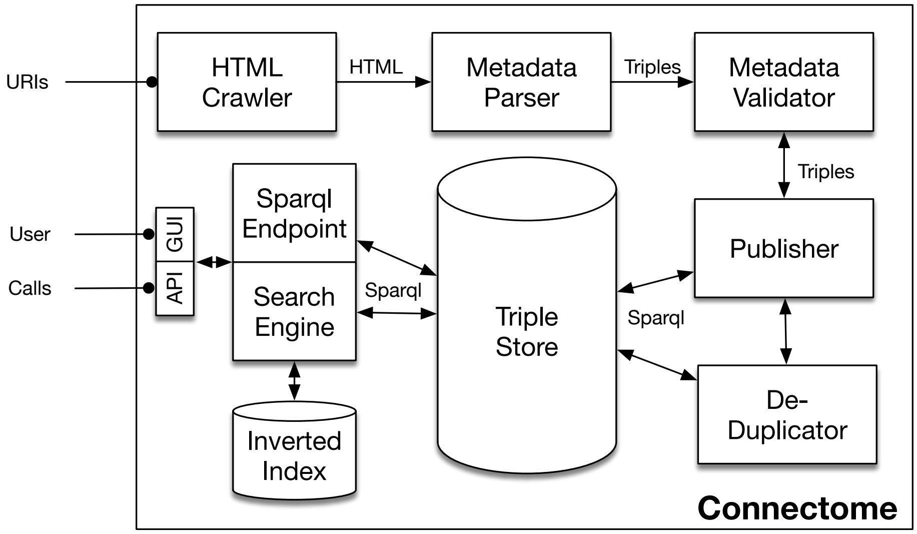


Figure 1: A high-level architecture of the Connectome MVP

- The **HTML Crawler** takes as input subdomains and crawls webpages recursively;
- The **Metadata Parser** receives webpages from the Crawler and extracts embedded metadata (e.g., in Microdata, RDFa, or JSON-LD);
- The **Metadata Validator** takes the extracted metadata and tries to align them to the schema and format adopted by the Connectome; cleaning and validation processes (e.g., using SHACL) will have to be implemented in that context;
- The **Publisher** takes the resulting triples from the Validator and coordinates the publishing process of the new datasets;
- The **De-Duplicator** determines whether the dataset being published is new, is an update or a fork of an existing dataset known by the Connectome;
- The **TripleStore** is the main data repository of the Connectome and store all metadata related to the system;
- The **Sparql Endpoint** receives, validates and processes SPARQL queries received from the outside;
- The **Search Engine** receives, validates and processes keyword and faceted search calls using the Inverted Index and the TripleStore;
- The **InvertedIndex** duplicates parts of the metadata to enable efficient lookup queries (e.g., to answer keyword queries);

- The **API/GUI** layer, finally, provides interfaces to access the Connectome (both for humans and machines).

This MVP offers the core functionalities discussed above. It can also be extended in a number of ways described below.

4.3 Extension for Handling Older Datasets

Handling older datasets is possible, by developing dedicated wrappers responsible to align the metadata from existing standards or repositories to the Connectome's schema. Results will vary, as they depend mostly on the available metadata that are available to describe older data (format, quality, coverage of the metadata).

4.4 Extension for Indexing & Connecting Raw Data

Indexing, and potentially connecting, raw (instance) data is desirable to power more effective queries (i.e., queries on the data rather than on the metadata only). However, similar to the point above, a new component will have to be developed for each format one ones to support. The component should be able to ingest the raw data, and to output corresponding triples (for the triple store) and key-value pairs (for the inverted index) to index the raw data. The additional overhead in terms of processing and storage can be high in that case—depending on the format and the size of the raw data.

4.5 Extensions for Provenance

As noted above, advanced provenance capabilities could be supported by working hand-in-hand with third-party tools (e.g., RENKU). Advanced provenance queries could then be supported (e.g., "give me the latest version of this dataset" or "give me all datasets derived from this original data") by storing and handling the corresponding metadata (e.g., in PROV-O), and by offering new query modalities (either for humans or machines). Provenance could also be leveraged to improve search results and de-deduplication in our context.

5 Governance

Finally, I explore below a few ideas related to the establishment and proper operation of the Connectome.

5.1 Schema & Functionalities

Similar to what schema.org is doing, I propose to discuss the high-level functionalities (what should the Connectome do) as well as the schemas supported by the Connectome by two entities:

- An open **Community Group** discussing functionalities and potential improvements to the Connectome, potentially through an online forum and an issue tracker (e.g., Github);
- A smaller **Steering Group** meeting regularly to accept changes and decide on a schedule for the releases of the schema and the software. Decisions regarding the schemas could also be outsourced to existing entities (e.g., W3C or schema.org). It is important that prominent Swiss research entities are represented in this Steering Group, as the Connectome cannot function without them (e.g., to provide guidelines or incentives to the researchers to publish data).

5.2 Software

The software powering the Connectome should be open-source, and should probably be maintained by one entity or a small set of entities working tightly-together. Parts of the architecture could be based on existing solutions, as mentioned above in Section 2.3, though one should proceed with care in that regard as integrating heterogeneous pieces of software is always very delicate.

The software should run on state-of-the-art infrastructures (e.g., containers on a cloud infrastructure) hosted in Switzerland.

Funding options for developing this software are open, but would ideally be granted from Federal sources to enable long-term deployment and operation.

The Steering Group should also probably publish and maintain a list of preferred solutions to interact with the Connectome (e.g., solutions to publish datasets, or to create webpages containing metadata describing datasets, etc.).

5.3 Legal Considerations

Legal, privacy and transparency issues should be carefully checked with an expert. The resulting platform should if possible be considered as *non-commercial*, as defined by the SNF in its Data Management Plan guidelines[39]

In addition, the platform should support the recommendation, creation and publication of licences accompanying the data. This point should be investigated by a legal expert, although Creative Commons licenses[40] should probably be preferred as they are already widely adopted for creative works online.

6 Conclusions

Making research data more findable, accessible, interoperable and reusable offers many key advantages, such as accelerating research, promoting interdisciplinary collaborations, promoting reproducibility or creating more powerful, data-driven

[39] http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx

[40] https://wiki.creativecommons.org/wiki/data

(e.g., Machine Learning) models. In light of the current best-practices, standards and systems reviewed in this document, I believe that the time is ripe for establishing solutions in that context. With SWITCH's vision for a research data Connectome, Switzerland could play a leading role in that field by offering a ready-to-use platform to our scientists. I hope that this document can pave the way for establishing such a platform by providing a pragmatic but powerful solution in that context.

References

[1] Maxi Kindling, Heinz Pampel, Stephanie van de Sandt, Jessika Rücknagel, Paul Vierkant, Gabriele Kloska, Michael Witt, Peter Schirmbacher, Roland Bertelmann, and Frank Scholze. The landscape of research data repositories in 2015. *D-Lib Magazine*, 2017.

[2] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. The trials and tribulations of working with structured data: -a study on information seeking behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1277–1289. ACM, New York, NY, USA, 2017. ISBN 978-1-4503-4655-9. URL `http://doi.acm.org/10.1145/3025453.3025838`.

[3] Xin Luna Dong and Divesh Srivastava. *Big Data Integration.* Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015. URL `https://doi.org/10.2200/S00578ED1V01Y201404DTM040`.

[4] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[5] Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Jeffrey S. Grethe, Hua Xu, Ian M. Fore, Jared Lyle, Anupama E. Gururaj, Xiaoling Chen, Hyeon-eui Kim, Nansu Zong, Yueling Li, Ruiling Liu, I. Burak Ozyurt, and Lucila Ohno-Machado. Dats, the data tag suite to enable discoverability of datasets. *Scientific Data*, 4(1):170059, 2017. ISBN 2052-4463. URL `https://doi.org/10.1038/sdata.2017.59`.

[6] Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, and Tim Clark. A data citation roadmap for scientific publishers. *Scientific Data*, 5(1):180259, 2018. ISBN 2052-4463. URL `https://doi.org/10.1038/sdata.2018.259`.

[7] Marcin Wylot, Philippe Cudré-Mauroux, Manfred Hauswirth, and Paul T. Groth. Storing, tracking, and querying provenance in linked data. *IEEE*

Trans. Knowl. Data Eng., 29(8):1751–1764, 2017. URL https://doi.org/10.1109/TKDE.2017.2690299.