SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

D6.1 INVENTORY AND ANALYSIS OF DATAFLOWS

SHARIF ISLAM, TINA LOO, WOUTER ADDINK, HELEN HARDY

DOI: 10.5281/ZENODO.4085027

Grant Agreement Number | 823827
Acronym | SYNTHESYS PLUS
Call | H2020-INFRAIA-2018-2020
Start date | 01/02/2019
Duration | 48 months
Work Package | WP6
Work Package Lead | Wouter Addink
Delivery date | 31.07.2020 – extended to 31.10.2020

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

Contents

Table of Contents

Summary

The European Loans and Visits System (ELViS), a deliverable of SYNTHESYS+ WP6, will implement a one-stop shop for researchers to provide open access to over 490 million specimens at 21 institutions. ELViS will support workflows for Virtual Access (VA) and Transnational Access (TA) calls and for individual loans, visits, and digitization requests. In order to facilitate these workflows, ELViS will use, collect, integrate, and generate data that will flow through different components. This deliverable (D6.1) identifies these relevant data sources and the associated data flows. The key objective is to ensure that the source and nature of these data resources and flows is documented and available for future development initiatives facilitating collection access across institutions.

The DiSSCo Provisional Data Management Plan outlines two types of data, along with their meta-data, within the context of new services that will be developed. The primary category is associated with digital specimens and collections such as specimen and collection data, annotations, interpretations. In addition, multiple secondary categories of data (such as details of researchers and collectors, loans and visits requests, equipment and laboratory

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

details) are expected to be managed and used by DiSSCo. The data in ELViS fall into the secondary category and are closely linked to and dependent upon the primary category. Within this framework, we further categorise the data sources: external and internal. External data sources implies that the owner and maintainer are outside of the core DiSSCo institutions (for example data from CETAF and GBIF). Internal data sources are generated within ELViS and DiSSCo (for example, loan and visit requests data). These data sources in ELViS are part of specific data flows that make up different components of the system. The data flows are sequentially envisioned as such: 1) Authentication and authorization 2) Request: Transnational and Virtual access, loans, visits, digitization on demand and 3) Report.

It is vital to understand where the data are coming from and the nature of the data in order to provide input for requirement analysis and feature development. This report of data inventory and flow analysis thus provides valuable input for system and service development, gap analysis, risk assessment, and improving existing workflows so ELViS can provide better access to the collections.

Description of Deliverable

The deliverable is a report that includes the context of why different data sources are needed for ELViS. Then it provides a list of different types of data sources in a table highlighting the relevant properties. The report proceeds with analysis and examples of data flows grouped in three major categories: i) Authentication and Authorisation; ii) Request ; iii) Report. ELViS is still in the early stages of development so some of these data flows will be refined further as insights evolve and mature.

Background

The ELViS system will facilitate the placement, assessment, prioritisation, and monitoring of requests for visits, loans, and digitization. Hence it will primarily be a transactional system that needs to capture all communications and decisions around these requests. In order for these transactions to be embedded in the day to day workflow, the system needs access to detailed information about the collection holding institutes which includes but is not limited to: institutional profile, facilities information, researcher and expertise profile, collection descriptions, specimen details and collection status. It will also need a system for discovering and linking these datasets. To avoid duplication and promote efficiency ELViS will need to make use of already existing data resources trusted by the community rather than creating new data platforms. Therefore a clear understanding of these data sources is crucial for system and service development, gap analysis, risk assessment, and improving existing workflows.

Scope

ELViS as a service will be primarily handling requests such as loans and visits and hence the primary scope of this report focuses on request-related data flows (see Figure 1 to see ELViS in the context of various other DiSSCo services). However, in order for these workflows to be effective, external and internal data need to be harvested, generated, and linked.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Therefore this deliverable looks beyond the service scope of ELViS to external data sources and services within the context of ELViS functionalities and related workflows.

In collecting this information the following was considered:

1. Which available data sources are required for ELViS functionalities?
2. Which data sources are required for the current SYNTHESYS+ requirements?

Not in scope: Workflow details, technical descriptions of the data linking and harvesting, description of the roles and actors. These will be addressed in other deliverables.

Methodology

Within the context of WP6 we have identified different processes in ELViS for the first phase of ELViS development efforts. Here are the list of activities and resources that provided input for the data inventory and flow analysis:

- Requirement Analysis for the DiSSCo Research Infrastructure (Raes et al. 2019)
- A survey (June 2019) to gather user needs which resulted in ELViS User Stories
- JRA1 workshop on ELViS System Design (MS47, July 2019)
- JRA1 weekly meetings
- Workshops and discussions during BiodiversityNext 2019, Leiden, the Netherlands.
- Plan for CETAF Collections Registry
- GeoCase (also see Petersen, Hoffmann and Glöckler 2019)

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

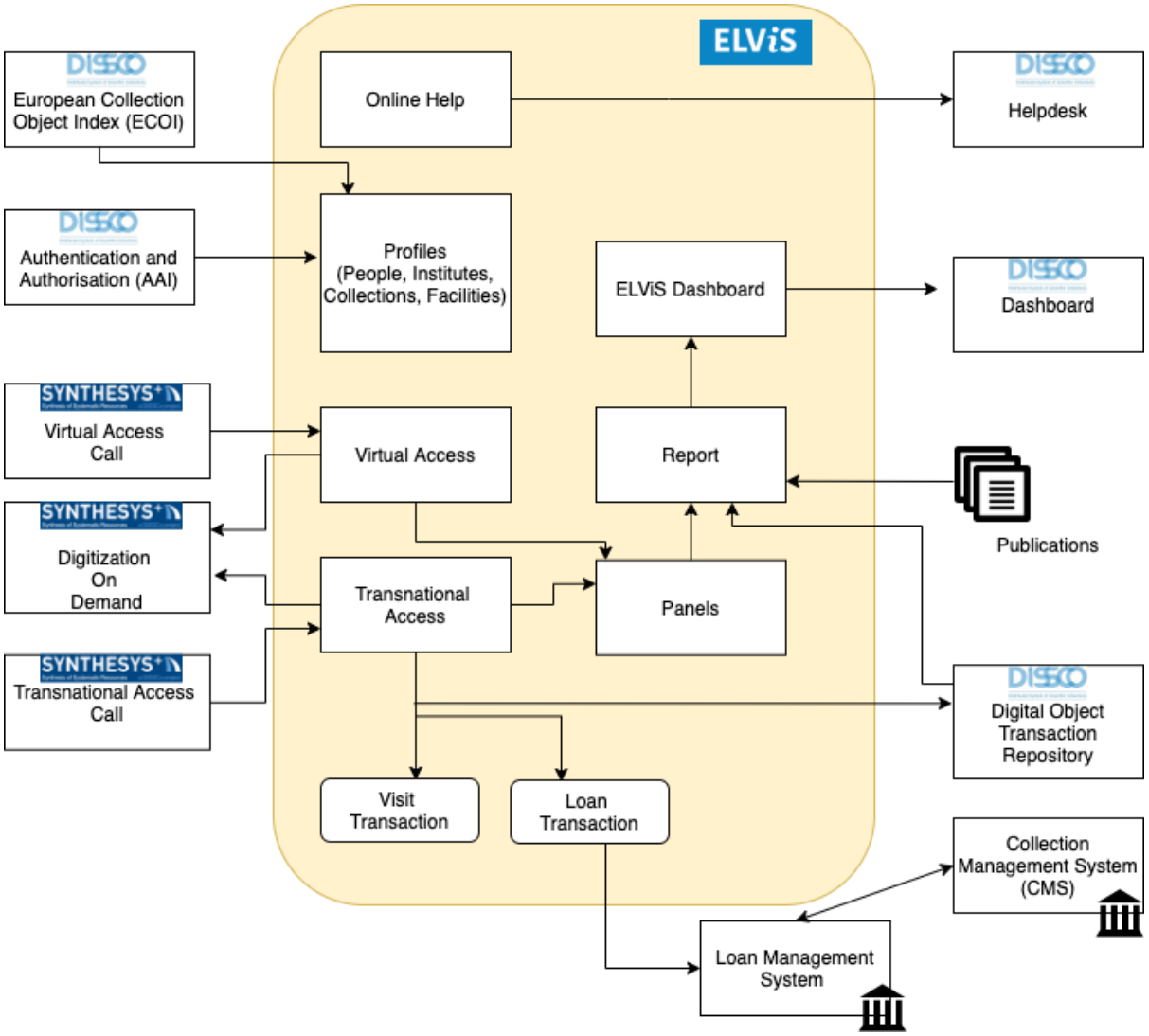

Figure 1: A data component view of ELViS within the context of DiSSCo services.

Data Inventory

The DiSSCo Provisional Data Management Plan (DiSSCo DMP 2019) outlines two types of data. The principal category is data associated with digital specimens and collections such as specimen and collection data, annotations, interpretations. In addition, multiple secondary categories of data are expected to be managed by DiSSCo. The data within ELViS falls into the secondary categories.

Below is a high-level diagram that visualises the data flow within the context of ELViS workflows. Here ELViS is a DiSSCo service that is designed to support specific requirements for loans and visits requests. There are different types of internal data storages and caches where these requests will reside. The data source can have both uni- and bi-directional connections with core DiSSCo FAIR Digital Object (Lannom, Koureas and Hardisty 2020) repositories and other external data sources.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

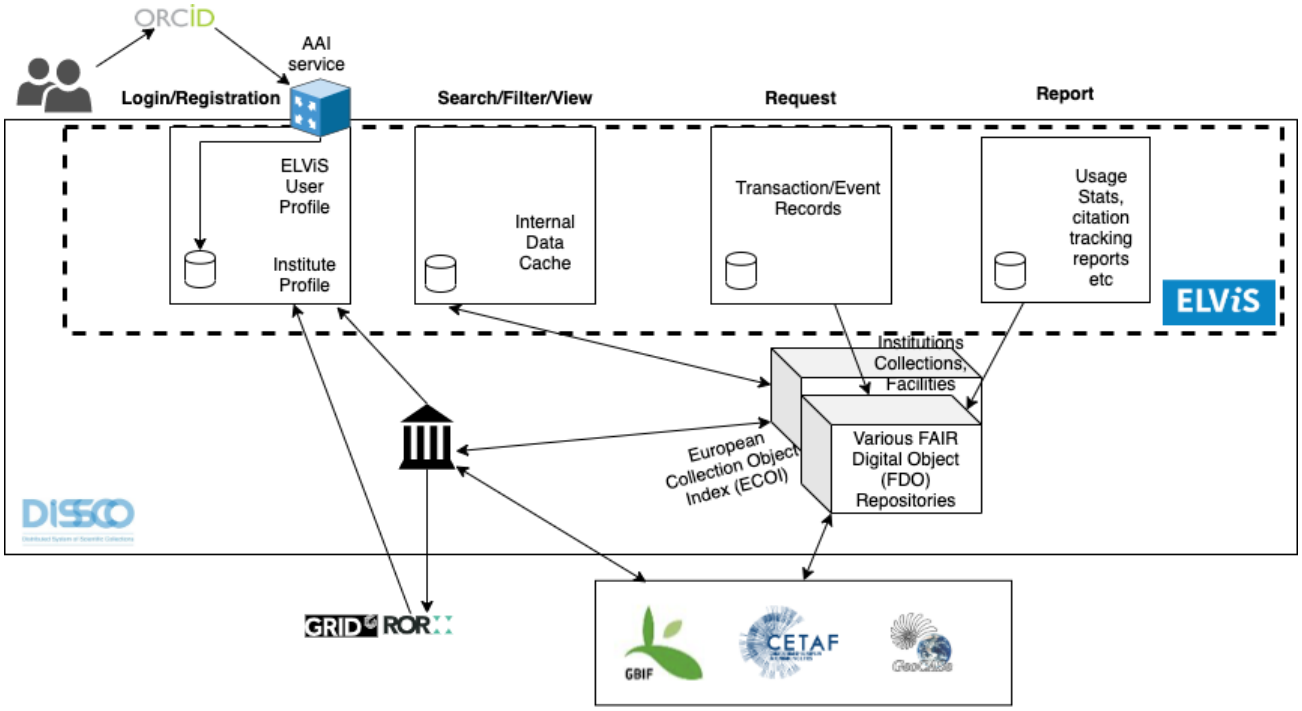

Figure 2: High level overview of ELViS data flow

The high level overview can be further broken into two categories: external and internal. External data sources are where the owner and maintainer are outside of the core DiSSCo services (in Figure 2: GBIF, CETAF, and GeoCase). Internal data sources are generated within ELViS and DiSSCo services (for example, the transaction records).

External data

ELViS needs authoritative data about institutions, people, collections and specimens from sources outside of DiSSCo core services. These are the criteria that ELViS will use to define whether a data source is authoritative:

- Data are curated by experts
- Data are provided directly by the source (e.g. collection data is provided by the collection holding institutes)
- Data are kept up to date
- Adheres to DiSSCo described standards (MICS, MIDS, openDS – see glossary for details).

ELViS will also make use of data sources that are semi-authoritative: data curated by a wider public and not always directly provided by the source (such as wikidata). See Table 1 for a detailed list of external data sources.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Table 1: External Data Sources

Providing system/ Data Owner	Location of the Data System	Used as an authoritative source for	Alternative source	Data Update Policy and Frequency	Has API	Format
Catalogue of Life	https://data.catalogue.life	Taxonomic data		Monthly and Annual checklist	Yes	JSON, XML
CETAF Passports	CETAF Systems	Collection description, collection holdings and facilities	Data harvested directly from the institutes. European Collection Object Index (ECOI)	Periodic data quality check on behalf of CETAF	No, one-time import	Excel, JSON
CETAF Collection registry	CETAF Systems (this is a new service currently under development and will replace the CETAF Passports data input and storage layers)	Description of collection holdings and facilities	Data harvested directly from the institutes. European Collection Object Index (ECOI)	Users can update data on an adhoc basis; periodic data quality checks will be implemented	Will provide API in the future	Excel, JSON
DataCite	https://datacite.org/	Identifiers for published datasets		every eight hours		JSON, XML
GBIF registry (formerly GrSciColl)	https://registry.gbif.org/ (the new registry is still under development)	Global Collection identifiers	CETAF Collection Registry	GBIF data crawler can update datasets as they become available	Yes	JSON
GeoCASe	geocase.eu	Digitised geological specimen		TBD	No, needs a one-time import until an API is available	

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

GGBN	http://data.ggbn.org/ggbn_portal/	May provide additional information about tissue collections		Similar to GBIF data pipeline		JSON
GRID	grid.org	Global identifiers for collection holding institutes	ROR (Research Organization Registry)	Updated 4 times a year (see policies: https://www.grid.ac/pages/policies)	Yes	JSON
Index Herbariorum	http://sweetgum.nybg.org/science/ih/	Botanical collection information	GrSciColl	Manual update	Yes	
ORCID	orcid.org	Global identifiers for users	Institutional Identity provider service	https://orcid.org/about/what-is-orcid/policies	Yes	JSON
SYNTHESYS portal	https://www.synthesys.info/	Data about facilities at institutes that may be available for use		Periodic update		Excel
WikiData	https://www.wikidata.org/wiki/Wikidata:Main_Page	May provide additional information about institutes or collections		Manual update	Yes	

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Internal data

Internal data are generated within the specific service contexts of ELViS and DiSSCo. For example, data related to handling loans, visits, and digitization requests. These data will be linked various types of digital objects (specimens, collections) in the DiSSCo infrastructure. See Table 2 for a detailed list.

Data Flow

Based on ELViS requirements and functionalities, data flows can be categorised in the following manner:

1. Authentication and authorisation
2. Requests (Transnational and virtual access, digitization on demand, loans, visits)
3. Report (statistics, reports, citation and research output tracking)

Authentication and authorisation data flow

Each ELViS workflow involves people in specific roles. ELViS manages the role by providing permission for certain actions to be performed on certain data in the system, however, the person in the role also needs to be authorised via an authentication and authorisation infrastructure (AAI). At the moment ELViS is planning to use ORCID as an external source for authentication (data elements such as email and affiliations can be stored in ELViS). In WP6, D6.2 (Piloting Access through an AAI infrastructure) we will pilot AAI infrastructure that will allow users to make use of federated authentication (users will be able to use their own institution's email and login credentials to access ELViS).

Example data flow: Virtual Access (VA) coordinator logs in using ORCID or the institutional credential to ELViS and views proposals written by collaborators from three different institutions. With AAI data flow, all the collaborators are identified and authorised via their own institutional credentials. Data flow here is going from institutional or federated identity providers to ELViS.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Table 2: Data generation within the ELViS and DiSSCo system

Providing system/ Data Owner	Location of the Data System	Used as an authoritative source for	Alternative source	Data Update Policy and Frequency	Has API	Format
Digital Object	DiSSCo FAIR Digital Object repositories	Various types of Digital objects in DiSSCo (Collection Descriptions, Digital Specimen).		TBD	Yes (work in progress)	JSON, JSON-LD
ELViS	https://elvis.dissco.eu	Loans, visits, digitisation requests, usage reports.	Institution's loans and visits database (if structured and reliable data exists)	TBD	Yes (work in progress)	JSON

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Request Data flow

The main functionality of ELViS is to facilitate requests related to collection and facility access. A request data flow involves allowing the requester to submit the request; registering and acknowledging the request; notifying related parties; allowing the request handlers to handle the request and provide responses; closing/updating the request when done. It is also important to note that institutions currently and will in the future use several different types of specialized software (such as content management systems). Even though discussions around standards are happening in different venues (e.g., SYNTHESYS+ NA4, DiSSCo Prepare WP5, TDWG Collection Description Interests Group, DINA consortium) and ELViS is working towards creating a one-stop shop, the workflows cannot be fully handled in a single system. Therefore ELViS components need to integrate and interoperate with other systems. Task 6.3 (Testing and integration of workflows) will explore these integration, compatibility and interoperability issues. In the future, due to the deployment of digitisation on demand and other DiSSCo e-services, different workflows will emerge and the nature of loan requests will evolve (e.g., a digital request might be handled first before a physical loan request after assessing the need based on the research question).

Example loan request data flow (see Figure 3, different roles in ELViS are indicated by different colors): A requester interested in a specific object (searchable in the European Collection Object Index) requests it as a loan. The requester may have queries about the object or the collection. ELViS in that case redirects the question to the appropriate subject expert. The status of the request is also provided to the requester. Request handler from the collections-holding institute reviews the request and makes a decision and updates the status of the item in the ELViS system accordingly. In certain cases, some of these loan requests can turn into digitization requests. All the transactions are stored in the event log which is fed into other workflows such as the generating reports. The loan data can also flow to external systems maintained by the institute.

Report data flow

As the requests are stored, transactions reports can be generated depending on the stakeholder needs. Example data flow: a curator generates a report about loans and visits within a particular date range for a report that will be published in a journal or on the institutional website.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

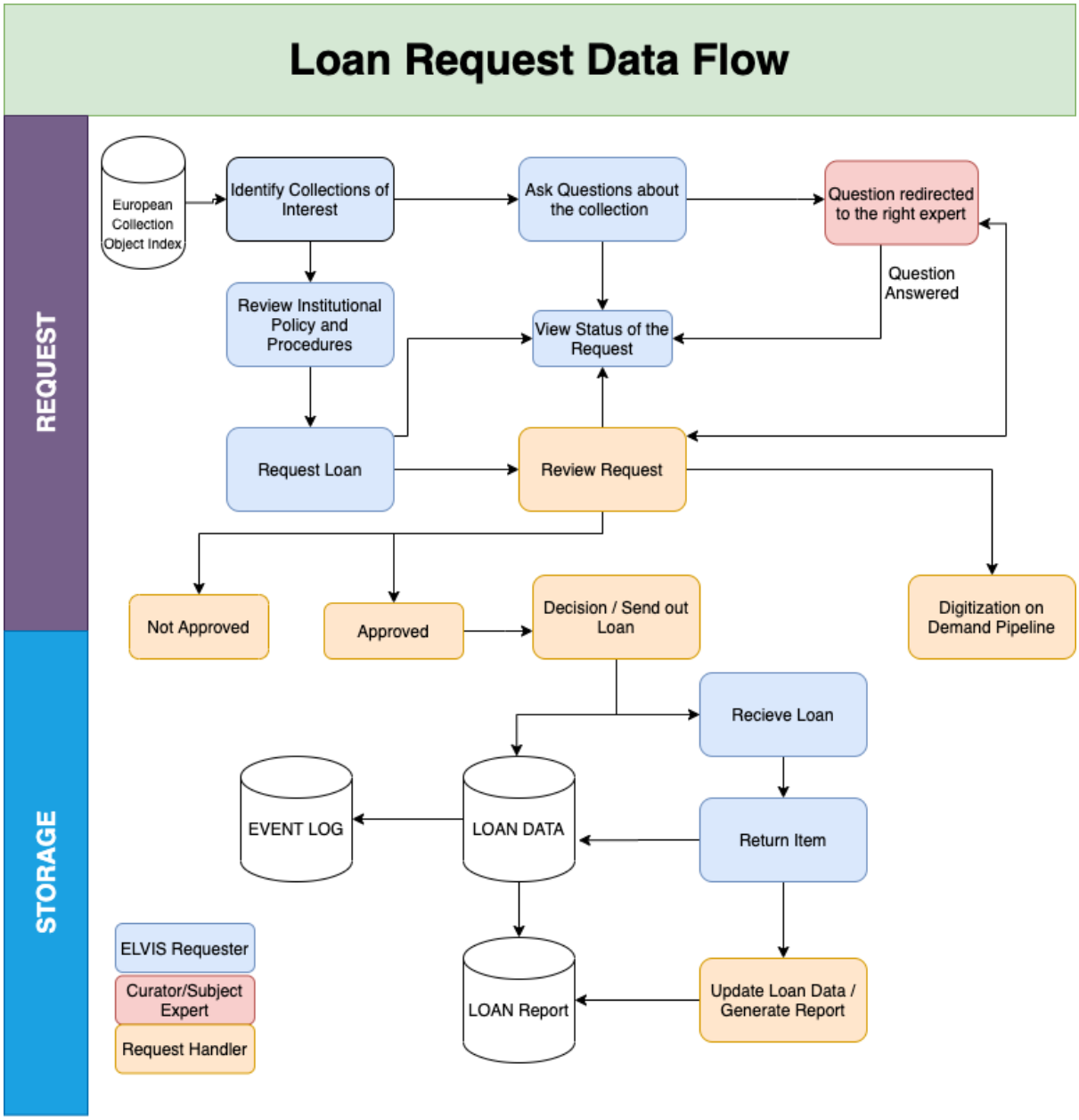

Figure 3: Example Loan Request data flow visualisation

Future development

As ELViS will be interacting with a variety of different data sources, it is crucial to have an overview of external and internal data sources. This report will facilitate further planning of ELViS development by providing detailed insights into the types of data sources and flows. During the requirement analysis, design, and development phase, these flows will be further refined and expanded as the landscape changes.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Glossary of Terms

- JRA1 – ELViS is a project under the Joint Research Activities stream JRA1: Optimisation of Access (Smith et al. 2019).
- TA - Transnational Access. SYNTHESYS Project funding is available to provide scientists to undertake short visits to utilise a Taxonomic Access Facility (TAF).
- VA – Virtual Access. The Virtual Access is a new step for the SYNTHESYS programme. It gives researchers an opportunity to propose and make the case for a collection or collection item(s) to be digitised by the holding institution for the wider benefit of the collections community.
- MIDS - Minimum Information about a Digital Specimen.
- MICS - Minimum Information about a Collection (DiSSCo DMP 2019).
- openDS - open Digital Specimens is a specification of Digital Specimen and other related object type definitions essential to mass digitisation of natural science collections and their digital use in a new generation of infrastructure and applications (Hardisty et al. 2019).

References

Hardisty, A., Ma, K., Nelson, G. and Fortes, J. 2019. 'openDS'–A New Standard for Digital Specimens and Other Natural Science Digital Object Types. Biodiversity Information Science and Standards, 3, p.e37033.

DiSSCo DMP. 2019. Provisional Data Management Plan for the DiSSCo infrastructure. DOI: https://doi.org/10.5281/zenodo.3532937

Lannom, L., Koureas, D. and Hardisty, A.R., 2020. FAIR data and services in biodiversity science and geoscience. Data Intelligence, 2(1-2), pp.122-130. DOI: https://doi.org/10.1162/dint_a_00034

Petersen M, Hoffmann J, Glöckler F. 2019. Access to Geosciences – Ways and Means to share and publish collection data. Research Ideas and Outcomes 5: e32987. DOI: https://doi.org/10.3897/rio.5.e32987

Raes, N., van Egmond, E., Addink, W. and Hardisty, A., 2019. Requirement Analysis for the DiSSCo Research Infrastructure. Biodiversity Information Science and Standards. DOI: https://doi.org/10.3897/biss.3.37892

Smith, V., et al. 2019. SYNTHESYS+ Abridged Grant Proposal. *Research Ideas and Outcomes*, *5*, e46404. DOI: https://doi.org/10.3897/rio.5.e46404

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.