EUDAT

EUDAT Community Engagement

Core communities and Data Pilots

EUDAT works directly with a wide range of research communities to deliver common data services to support and resolve their research data management challenges. To be successful in this ambitious initiative, EUDAT uses novel methods to involve all the stakeholders, both in the discussions to determine the required services, and in the process of designing, developing and implementing those services. These methods include involving communities in the core Research, Innovation & Development activities, known as EUDAT Core Communities, as well as collaborating with communities through specific data Pilots. This booklet gives an overview of EUDAT's 7 core communities and 24 data Pilots currently running. For more information see www.eudat.eu

Social Sciences and Humanities

There is a one 1 community member and 5 data pilots represented for the Social Sciences and Humanities

Core Community - CLARIN

The CLARIN project is a large-scale pan-European collaborative effort aimed at making language resources and technology readily available for the whole European Humanities (and Social Sciences) community. This includes coordinating the development of appropriate resources. Amongst other things, CLARIN will offer scholars tools for computer-aided language processing. In more detail, CLARIN offers:

- Comprehensive services to the humanities disciplines with respect to language resources and technology.
- Technology for overcoming the many barriers created by institutional, structural and semantic interoperability problems and fragmenting the resources and tools landscape.
- Tools and resources that will be interoperable across languages and domains, thus addressing the issue of preserving and supporting the multilingual and multicultural European heritage.
- Comprehensive training and education programs that include university education in the different member states.
- Improvement and extension of web-based collaborations, i.e. creating virtual working groups breaking the discipline boundaries.
- Development or improvement of standards for language resource maintenance.
- A persistent and stable infrastructure that researchers can rely on decades to come.

Areas of collaboration with EUDAT

CLARIN has been one of EUDAT's core communities since 2011 and the service that it has been most involved with, up to now, is B2SAFE. Various CLARIN centres are using it to perform safe replication of the language data they are hosting. The University of Tübingen (Eberhard Karls Universität Tübingen), the LINDAT/CLARIN Centre for Language Research Infrastructure in the Czech Republic (usually known as LINDAT/CLARIN) and the Max Planck Institute for Psycholinguistics in the Netherlands are all using B2SAFE. There are a further eight centres ready to use B2SAFE and tailored uptake plans are under development to be deployed over the coming months. CLARIN will harvest the B2SHARE metadata related to language material and make it accessible via their search portal, the Virtual Language Observatory. Additionally, EUDAT's B2DROP service has been tested and is used for internal data exchange and sharing. CLARIN community metadata has been integrated into B2FIND, EUDAT's metadata portal.

Main benefits for the community

EUDAT has a strong focus on facilitating the uptake of its services by research communities through specific uptake plans driven by the active involvement of community experts. This collaboration between EUDAT & CLARIN makes it possible for "homeless" researchers and

citizen scientists to deposit their resources into a good data repository with long-term preservation. There are other services – for example, the metadata integration into the infrastructure, or safe replication – that are all the kinds of services on which CLARIN and other research communities can build. It makes sense to pool these resources and make sure there are good, reliable and stable services that can be used by everyone as, in the end that will benefit all the research communities.

For more information on CLARIN see: https://www.clarin.eu/

Research data repository for students' own results

Overview of the pilot

In the Department of Physics of University of Helsinki, the masters level training of physicists; have traditionally included extensive laboratory experiments, their documentation and reporting. This pilot is for including data publication and curation of the experiment results in the laboratory courses: storing the observations, together with relevant metadata into a repository, where the course assistants would have access. The students would then learn to publish and document their data as a normal part of scientific workflow. Naturally it would be needed also to include methods to "cite" the data sets using a PID offered by the system.

The scientific & technical challenge

One of the biggest issues in data sharing is a cultural one. Current research paradigm does not necessarily consider data sharing and curation as a part of normal scientific process. Teaching this as a part of normal course work is the way to get the message across to new scientists, thus creating possibility of new generation of scientists who do not need to be taught and forced to publish their data – no more than they are to publish their results. Taking data publication as a normal part of course work is the key.

The data publication should however be realistic, easy and flexible. The system should be similar which are used in long-tail of the research data applications and the overall system should also serve the overall course work. For this reason, the system should be capable of authentication, team work, controlled evaluation and to be completely citable. The annual number of students in the initial phase is less than hundred and the data amounts are small.

Why EUDAT?

From the EUDAT side the system will be realised using a version of EUDAT B2SHARE platform, with some local variations (access rights, template). The access to results would (at least in default) be for the student (data originator) and the course assistants (for control). The students would have realistic, but practical user experience for publishing

small datasets. The overall size of the datasets is not large. Crucial issue is a good interface (including somewhat specified template) and ease of use. The template will be developed by UHEL together with EUDAT2020 team. There should be an unique identifier for each data set for realistic inclusion to the students' reports

Two optional additions can also be considered:

- In long run, the repeated experiments could be of interest to e.g. pedagogical research. Thus possibility of anonymization of data set producers and completely open access to results after a grace period could be optional goal.
- Another optional goal could be direct commenting of the data sets (in this case by course assistants), and if corrections are needed - new versions of data sets.

Expected outcomes

New physicists graduating from the Department of Physics will take data publication and curation as a natural part of their scientific work. This will increase their employability and significance of their research. The education will also give secondary benefits when part of these students will continue as PhD students in the research teams. The education of data publication and curation is also an important transferrable skill. This system can also be then generalized on different courses throughout the University of Helsinki and Finnish education sector – all important users of EUDAT services.

Expected domain legacy

Most important factor is to foster the culture of open research. Teaching students early on that data publication is crucial and natural part of the scientific process is one of the key ways to make this change. Experiences on metadata inclusion with from a large pool of students can be useful to find practical bottlenecks of such systems.

Enriching Europeana Newspapers

Overview of the pilot


Le Petit Parisien

C'EST UNE PAGE UNIQUE DES ANNALES DU CRIME

COMMENT BONNOT FUT PRIS ET TUÉ

BONNOT

DUBOIS

Enriching Europeana Newspapers aims to expose the full text aggregated as part of the Europeana Newspapers project. It contains over 11 million pages of full text of historic newspapers (mainly but not all 19th century), drawn from national and research libraries across Europe. A portal is already in place at http://www.theeuropeanlibrary.org/tel4/newspapers. This pilot aims to expose and improve the text for more data driven usage (ie large scale data analysis of the whole corpus)

The scientific & technical challenge

The key scientific challenges are these:

- Creating best practice guidelines for the publication, citation and impact measurement of cultural heritage data (ie the newspapers in question). Standards for citing and judging the impact of open cultural data are still far from being established.
- Enriching the newspapers corpus, via the automatic extraction of topics and named entities; the current corpus is only searchable via free text searches
- Showcasing the value of the enrichment by a quantitative analysis of the occurrence of topics/entities over time and across borders.

A particular challenge will be the extraction of topics across texts in multiple languages (over 40 languages are featured in the corpus from French to Yiddish to Estonian) and variable quality of the digitised text.

Digital humanities scholars will be interested in the raw OCR texts; the number of these is likely to be in the 100s rather than 1000s. We also suspect that others in linguistics, economics, information science and computer science can make use of the datasets

If successful the enriched texts could also be placed in the current Newspapers interface (http://www.theeuropeanlibrary.org/tel4/newspapers). This received over 1.4m page impressions in 2015, around 5 to 6,000 users a month. Better search facilities will help improve these numbers

Why EUDAT?

We will be using the B2SAFE and B2FIND services. These will help us undertake the enrichment of the datasets and, more generally, expose them for re-use by other academics,

particularly those outside the digital humanities. At present, users of the service tend to be 'traditional' historians who are familiar with the search and browse possibilities of the portal – connecting with the tools and, just as importantly, the EU-DAT community.

Expected outcomes

We expect to meet to three significant use cases:

1. to have a better understanding of the topics, themes and subjects featured in the newspapers, allowing researchers richer understanding of the how certain issues and ideas were phrased in the corpus under question
2. the extraction of topics will also assist with resource discovery – allowing users of The European Library portal to search not just for free text words but topics. (Note: the re-integration of the enriched dataset into the existing newspapers portal is not foreseen in this piece of work)
3. exposure of the datasets via EUDAT will allow for much greater discovery and reuse of the newspapers corpus as a whole

Expected domain legacy

Meeting the use cases described above will help the study and understanding of historic newspapers as source material. The use of news[papers has been standard within many disciplines in the humanities and social sciences for a while, but their availability as a 'big data corpus' opens up many methodological avenues currently unexplored. It enables new research questions on language, communication as well as any topic featured in the newspapers.

Creating and exposing a corpus drawn from so many different countries also has benefits in developing transnational history, ie exploring themes and relationships between different European countries or the continent as a whole .

Cloudy Culture: A study of EUDAT shared services to measure the potential of using cloud-like services to improve the preservation of digital cultural heritage

Overview of the pilot

For the benefit of cultural organisations the National Library of Scotland, working with Edinburgh Parallel Computing Centre (EPCC) and with the support of the National Galleries of Scotland and the Digital Preservation Coalition will explore the potential of EUDAT cloud-like services to preserve European digital cultural heritage. The pilot will inform practitioners in digital preservation, curation and archiving and will test 3 elements of the EUDAT platform:

- The online transfer of large amounts of data to EUDAT (c 100TB/20 million files)
- Safe storage of data over time
- The use of high performance computing to accelerate preservation actions

The scientific & technical challenge

Cultural organisations need to preserve access to an increasing amount of digital content that they are creating and acquiring. For example the National Library of Scotland expects its data to grow 10 times over 10 years. This growth increases the strain on the core preservation requirements to store data in multiple geographic locations (cost of setting up more data centres), and to check if data changes over time (costs of increased computing power/time). High level studies suggest that traditional cloud services offer no net benefit for large volumes of data (100s of TB) that require on-going access to undertake preservation actions. There is little openly published information that describes or quantifies the practical limits and costs of using cloud-like services. For example how long will it take to transfer data? Is transfer and data monitoring scalable? What additional tools and services are required to automate the process?

Cloudy Culture will use EUDAT to hold a safe preservation copy of data to allow locally held access copies to be repaired if they change over time. For this reason access to the copy at EUDAT will be restricted to those few people who are undertaking preservation actions on the data. However the local copy of the data, mainly digitised collections, is freely and openly available via www.nls.uk where the audience size is millions of visitor sessions per year.

Why EUDAT?

Running in parallel with the growth of digital cultural heritage is the development of large data centres with a focus on science data. The European Commission sponsored 2014 Digital Cultural Heritage Roadmap for Preservation identifies existing e-Infrastructures as a solution to this problem, connecting these facilities with digital cultural heritage to ensure

our heritage remains accessible and usable long term. The Cloudy Culture pilot, supported by EUDAT, wishes to exploit this same synergy.

Cloudy Culture will integrate local tools with those developed by EUDAT and use B2SAFE, B2STAGE, iRODS, storage and computing power at the EUDAT facility at the Edinburgh Parallel Computing Centre (EPCC) to:

- automate the managed transfer of digital content and metadata between the National Library of Scotland and EUDAT
- automate and report on preservation actions undertaken in the EUDAT environment such as fixity checking and file format characterisation, accelerated by the availability of large amounts of computing power

Expected outcomes

During the 18 month pilot the EUDAT services will enable a third copy of high value digital cultural heritage at the National Library of Scotland to be stored in a different location and on different technology than the other two copies kept locally. This reduces the risk of losing the same data from all copy locations and so improves preservation. Because the EUDAT copy is monitored for changes through fixity checking any unwanted changes can be identified and repaired using intact local copies and vice versa. In addition the Cloudy Culture team will:

- improve local and EUDAT tools and workflows and the automation of data transfer and preservation actions to reduce human resource requirements and make preservation more sustainable
- gain an improved understanding of using EUDAT services such as transfer and compute times, transfer stability, scalability and costs
- understand the potential to use EUDAT and other cloud-like services beyond the pilot phase

Expected domain legacy

Cloudy Culture partners, in particular the Digital Preservation Coalition, will help to disseminate the results of the pilot for the benefit of the wider digital preservation, curation, archiving and cultural heritage domains. By doing this Cloudy Culture will increase community understanding of using cloud-like services and improve the communities' decision making. Cloudy Culture will:

- share information about the costs and benefits of preservation storage using EUDAT that can be transposed to other cloud-like services
- understand the viability and limiting factors of cloud-like services for large amounts of data
- improve tools for automated data management and share these with accompanying documentation so they are useful to others
- expand EUDAT's potential to act as a digital preservation option for European digital cultural heritage

Aalto data repository

Overview of the pilot

We are creating a central online location for data sharing for all Aalto University researchers. This will host both data and metadata: the name, description, ownership, source, and information on usage. Other dataset hosting sites exist, so our main target use case expanding EUDAT scope is intra-Aalto University interaction. Researchers with data analysis skills will be able to find data related to their work, as well as the domain experts responsible for that data. Furthermore the solutions should be tightly integrated to existing computing resources and Big Data platforms available nationally.

The scientific & technical challenge

In Aalto University, Big Data and Data Science have been recognized as key areas of ICT and digitalization at all levels of rapidly developing socio-economic societies. These systems generate ever-increasing amounts digital data, which can in unprecedented ways serve as a gold mine for researcher of various disciplines to study as well as enable the private sector players and public sector to develop their services, processes and technologies. Hence there is need to respond and find solutions to this data deluge, which is also reflected in the Aalto University application for profiling of Finnish Universities.

For the solution we have identified a set of design requirements that may pose also as a technical challenge. A few such are the requirements of 1) the metadata of datasets is full text searchable, 2) published datasets are assigned a persistent identifiers, 3) there are no restrictions to the type of uploaded data, 4) the datasets in the system can be made public for all the world to see, 5) the tool imposes no restrictions to the type of research data stored, 6) the metadata templates offered by the system satisfy the needs of different fields of science and also national requirements, 7) the system should be integrated to the already existing user management system.

In the first phase the system will have tens of users. If found to be successful, the solution will be scaled first within the University and possibly, even beyond to a national level. The possible user base may increase tenfold or even further in these cases. As Aalto users are working with large data sets the data volumes can already in the pilot phase potentially extend to tens of terabytes.

Why EUDAT?

We engaged into discussion with Centre for Scientific computing (CSC) that coordinates EUDAT operations in Finland. Based on these discussions we have assessed especially B2SHARE & B2DROP functionalities based on the requirements.

The results look quite promising and from this analysis we can see that the B2SHARE service would probably be the most suitable tool for addressing both our publishing and data management requirements. As a part of the pilot a suitable metadata templates can be provided

in B2SHARE and customized for Aalto University. Further, as federated authentication via B2ACCESS is supported in B2SHARE is should be possible to authenticate to the national identity service within EUDAT services. Also interfacing to national metadata and storage solutions is of interest to us in the EUDAT pilot.

Expected outcomes

The resulting data platform enables system and method level development, resulting in research innovations but it also opens up possibilities for educational and training purposes on Data Science and Big Data. In the initial phase there is a need to concentrate on data policy and repository practices for contributing to increased research effectiveness and generating wider goals of data sharing and open data in alignment with CSC's planned Big Data platform, but also to start widening and enhancing the skill base for data related research. We see that EUDAT solutions can paly a major role in supporting the implementation of the ambitious research data management goals set by Aalto University.

Expected domain legacy

This approach can enhance the system and process level understanding of any area of research, basic and applied, from science to engineering and humanities. It will strengthen multi and cross disciplinary research and step up product development, as well as facilitate novel innovations and services. At the same time it lowers disciplinary boundaries both in the public and private sectors, thus adding to innovativeness and new breakthroughs in R&D, and covering all areas of society and science from discovery and security to general competitiveness. The data repository will be initially used mainly for research purposes, but yet at the same time it will serve as a pilot and education platform for methodological development, training, and data repository technology.

Ancient OCR: Storing, Cataloguing, Relating, and Exposing OCR Objects from the Open Philology Project

Overview of the pilot

The Open Philology Project at the University of Leipzig has developed a modular, multi-threaded OCR pipeline to reach our goal of digitizing 100,000 books in the next three years. This pilot project gives us a way to store, catalogue, and expose the results of this pipeline, from original image to final OCR results. The users of the EUDAT system will be at the University of Leipzig and Tufts University (USA). The users of the data would be the same as those of the Perseus Digital Library, i.e., researchers and students in classical languages worldwide.

The scientific & technical challenge

The Open Philology Project produces many different data types according to many different data standards, all of which refer to the same real object, e.g., page images, HOCR data, XML text, syntactic treebank data, GIS data, etc. That is, the data that we publish should be thought of as a possibly ever-expanding vertical collection of objects. Perhaps our greatest challenge right now is to discover how we can get these different types of objects to interact not only within our own collections but also with collections outside of our own that might transform portions of our data or vice versa.

To do this, we need persistent identification of text and related data object

- that is stable throughout the creation, curation, publication, and post-publication lifecycle
- that can be leveraged easily across projects
- which can allow for a wide variety of PID schemes without requiring code changes for each one
- that is supported by institutional infrastructure
- and still allows for domain and project specific PID schemes

We also need a means to formalize and express details about the data types we are working with without coupling them tightly to the identifier schemes used to identify data objects that adhere to them. And, finally, we need support for multiple different models of collections (e.g., both horizontal and vertical) and a well-defined CRUD API for working with them.

The potential size of the user group is the same as that of the Perseus Digital Library, i.e., currently about 500,000 unique users

Why EUDAT?

We believe that the Data Types Registry and the PID Types API are both relevant to this use case. The Data Types Registry provides a data model for formally expressing data types, an API for CRUD and Query operations on a data type registry, and fulfillment of a dependency for the PID Types API. The PID Types API provides a conceptual model for a PID record and a CRUD API for interacting with PID records which can abstract the differences between PID implementations. We would like to begin to formalize the data types we are referencing here, possibly through use of the Data Types Registry. And we would like to explore use of the PID Types API to abstract the differences between different identifier types, as ultimately we expect this to be integrated with other components of the Perseus infrastructure, which uses a variety of different identifier types.

Expected outcomes

The concrete benefits of using EUDAT would be that it gives our data types domain neutrality, it improves our data management practices, it enhances the ability for our data to be reused by others because the data would be more clearly and thoughtfully expressed and because EUDAT provides a documented API for interacting with it. It would also give us the ability to scale more rapidly to support new PID types.

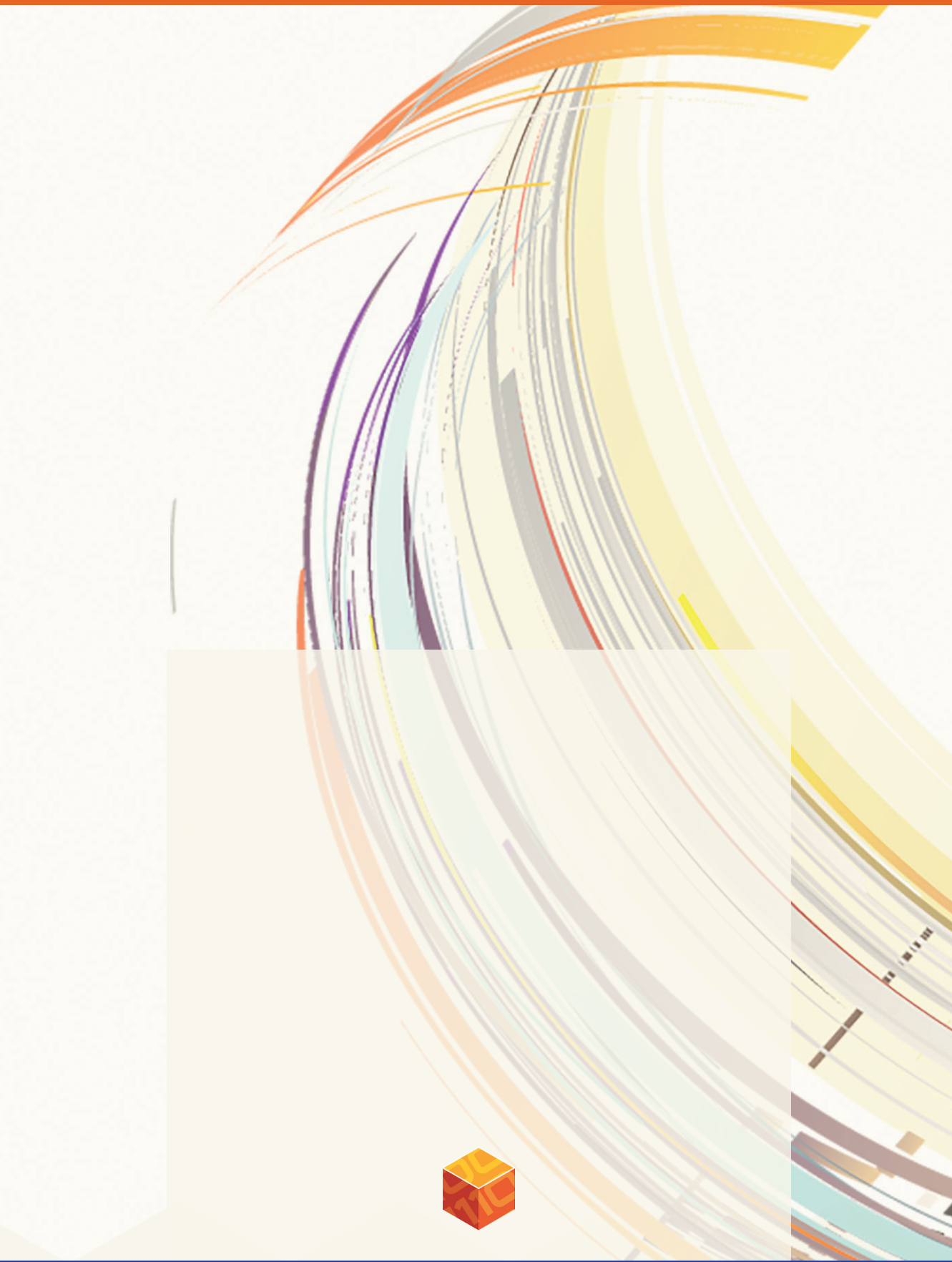
EUDAT could also provide the potential benefits of

- greater sustainability
  - if implementations of the APIs are built and maintained by diverse communities and projects
- greater bi-directional interoperability
  - if other projects with whom we want to share data find the same benefit and implementations
- Institutional support and long-term preservation
  - if our libraries and institutional repositories also deploy and implement the solutions

Expected domain legacy

Our goal at the Open Philology Project is to produce high quality, digital editions of every ancient work. EUDAT will help us to reach this goal by providing a way for us to store and reference our data, making it available to users on the outside. This availability will not only be to use the data but also to manipulate and improve it. In the realm of OCR, the best example of this will be the ability to access a page image and its resultant OCR results through the OCR proof reader we have designed, allowing them to check and correct the OCR results. Without EUDAT, we would be able to do this only on a very small scale. EUDAT will make it possible for any user of the Perseus Digital Library to engage in this vital activity of textual enhancement.


Apium emorroidarum.
genant sig blatern Eppich.
Das.ix.Capittel.
Fistulogia rotunda.
das ist Holwurtz.
Das.x.Capittel



European Commission | EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.