From disparate disciplines to unity in diversity

How the PARTHENOS project brings Humanities Research Infrastructures together

Authors: Frank Uiterwaal, Franco Niccolucci, Sheena Bassett, Steven Krauwer, Hella Hollander, Femmy Admiraal, Laurent Romary, George Bruseker, Carlo Meghini, Jennifer Edmond, Mark Hedges

Autobiographical note:

All authors of this paper work within different scientific institutions in Europe, from a historical archive to universities to Research Infrastructures. However, they all collaborate under the umbrella of project PARTHENOS, which stands for 'Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies'. PARTHENOS aims at strengthening the cohesion of research in the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, Archaeology and related fields.

Authors

Frank Uiterwaal – NIOD Institute for War, Holocaust and Genocide Studies – f.uiterwaal@niod.knaw.nl

Franco Niccolucci – PIN Scrl - Polo Universitario "Città di Prato" – franco.niccolucci@unifi.it

Sheena Bassett – PIN Scrl - Polo Universitario "Città di Prato" – sheena.giess@gmail.com

Steven Krauwer – CLARIN ERIC & Utrecht University – s.krauwer@uu.nl

Hella Hollander – DANS – hella.hollander@dans.knaw.nl

Femmy Admiraal – DANS – femmy.admiraal@dans.knaw.nl

Laurent Romary – INRIA & DARIAH – laurent.romary@inria.fr

George Bruseker – FORTH – bruseker@ics.forth.gr

Carlo Meghini – CNR-ISTI – carlo.meghini@isti.cnr.it

Jennifer Edmond – Trinity College Dublin & DARIAH – jennifer.edmond@tcd.ie

Abstract

Since the first ESFRI roadmap in 2006, multiple Humanities Research Infrastructures (RIs) have seen the light of day. At a disciplinary level, they have supported archaeologists (ARIADNE), linguists (CLARIN-ERIC), Holocaust researchers (EHRI), cultural heritage specialist (IPERION-CH) and others. These are a few examples to scratch the surface of the breadth of research communities which have benefited from large-scale European collaborative projects.

While RIs in each field have developed discipline-specific services over the years, common themes can also be distinguished. All Humanities RIs address, in varying degrees, questions around research data management, the use of standards and the desired interoperability of their data sets across disciplinary boundaries.

This paper sheds light on how a cluster project developed pooled services and shared solutions for its audience of humanities researchers, RI managers and policy makers. In a time where the convergence of existing infrastructure is becoming ever more important – with the construction of a European Open Science Cloud as an audacious, ultimate goal – we hope that our experiences inform future work and provide inspiration on how to exploit synergies in interdisciplinary, transnational, scientific cooperation.

Shared challenges in the humanities field

Innovative research is best served by a climate which enables interdisciplinary and transnational approaches. While this may almost sound like a truism, examples illustrate how important this is for humanities research questions. No matter whether a scholar studies Franconian languages, old Anglo-Saxon archaeological remains or World War I photography; to be able to successfully address research questions which stretch beyond national or disciplinary boundaries, one needs to be able to combine sources of knowledge located in different countries. At the same time, however, a significant amount of source material has been conceived – and is still, to some extent, confined – within national or disciplinary boundaries; be it transcripts of parliamentary debates, civil registration records or historical newspapers. More recently, the phenomena of proprietary data types and file formats constituted additional technical restrictions to the availability of information.

To address this situation, the European Strategy Forum for Research Infrastructures (ESFRI) was assigned the task to coordinate the integration of scientific knowledge and expertise at a European level, as a first step in the creation of what is referred to as the European Research Area.[1] In the first ESFRI roadmap, which was released in 2006, the challenge laid out for the Humanities was defined as follows: 'The present major task is (…) to create pan-European infrastructural systems that are needed by the social sciences and humanities to utilise the vast amount of data and information that already exist or should be generated in Europe'.[2]

Thirteen years later, it is fair to say that Humanities RIs have been prominent in the European research landscape and have only grown in size and relevance ever since this first strategic plan. They pool data and expertise, exchange knowledge, collaborate and, consequently, enable innovation in their respective fields. Between 2006 and now, several projects have been initiated, of which some are now established as legal entities. The first roadmap already mentioned the Research Infrastructures (RIs) DARIAH[3] and CLARIN[4], respectively supporting the Humanities at large and language related studies. The CENDARI project encouraged research in its two-pilot historic periods (Middle Ages and the First World War) and made archival descriptions from all over the continent available in one archival directory.[5] Following a similar approach, EHRI brought dispersed Holocaust sources together in its EHRI portal while also forming a 'human network', fostering cooperation among Holocaust researchers.[6] Archaeologists organised themselves in ARIADNE[7] while IPERION-CH opened up services and facilities to those focussing on the restoration and conservation of cultural heritage.[8]

While the success of Humanities RIs resulted in an unprecedented wealth of aggregated research assets, it was felt that the risk of the creation of separate research silos deserved attention. This was also demonstrated in a survey amongst 110 research institutions across Europe, which indicated that 'The major challenge of a collaborative and connective pan-European research programme will be to harmonise digital research practices by drawing together the numerous national and, increasingly, multilateral digital research initiatives'.[9] At the same time, it also became increasingly apparent that different disciplines were struggling with similar challenges. In all of the fields mentioned above, researchers have not found it easy to distinguish which policies apply to their research data, to choose the right standards for structuring and storing them, and to make sure that these data are interoperable with the work of others.

To develop solutions to both this 'digital data deluge' in quantitative terms and to these shared problems, the project PARTHENOS was conceived.[10] Since May 2015 it worked actively to address these problems by developing a suite of products and services for the study of history; for language-related studies; for archaeology, heritage & applied disciplines (including archivists, museum experts, preservation specialists, digital curators etc.) and, to a lesser degree, the social sciences. Examples of project output are a research scenario-guided tool to introduce humanities scholars to standards, a digital collaboration space entitled the PARTHENOS Virtual Research Environment and a set of guidelines to make research data reusable, based on the experience of PARTHENOS' underlying research communities.

This paper will provide insight in the challenges which underlie a cluster project when it tries to bring Humanities RIs together in the development of such shared solutions. In practice, RIs are often built either at a disciplinary level (such as CLARIN-ERIC, DARIAH etc.) or at a fundamental level (Géant, EGI, OpenAire), but opportunities for efficiencies and knowledge sharing at the meso-level were being missed out in this emergent landscape for some time.

After briefly introducing the project and its requirements-based approach, this paper will demonstrate how PARTHENOS has aimed for synergies around policies, standards, interoperability and training opportunities respectively. After addressing the dissemination challenges of a cluster project, a use case will finally shed light on how this all comes together.

What is PARTHENOS?

The project name PARTHENOS stands for Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies.[11] The project brought together major European integrating initiatives in the field. Besides the durable, member-funded infrastructures from the ESFRI roadmap, i.e. CLARIN and DARIAH, it included the projects

CENDARI, ARIADNE, EHRI and IPERION-CH. The collaboration took place within a so-called cluster scheme, introduced by the EU Horizon 2020 programme to foster the aggregation of RI projects. The basis of collaborating within such a cluster, is the search for common solutions to shared needs to the benefit of distinct but related research fields.

As a project, PARTHENOS' had a limited life-span. As the Horizon 2020 funding scheme provided resources for four years, PARTHENOS formally ended in October 2019 (after a six-month extension). Nevertheless, a sustainability plan was put in place to sustain project output, saving it for the future. Conceivably even more important, was the project's effect in terms of bringing various disciplines closer together in the process, facilitating future collaborations by encouraging the adoption of common standards and policies.

An approach based on the needs of communities

To ensure that the products and services that PARTHENOS has created would be fit for the purposes of different user groups, the project started with the collection of user requirements. As a first step, this is of course best practice and hardly revolutionary. The challenge PARTHENOS faced however, was specific to the nature of its integration efforts. In order to build bridges between Humanities disciplines, PARTHENOS' products and services were designed to cover the needs of all involved fields. In realising this goal, it was of vital importance to both focus on commonalities between the different Humanities disciplines, while also making the tools granular enough to cater to discipline-specific needs. A researcher in the field of linguistics will often require different information than a policy or decision maker, and a digital archaeology teacher's requirements might be very different from those of a technical specialist in metadata interoperability.

Consequently, the first step of the project was to draw up an aggregated inventory of user requirements. As a cluster project, PARTHENOS was in the unique position to collate earlier studies, many of them produced by previous or still ongoing projects in which PARTHENOS partners were – or are still – involved. This allowed for an overarching inventory of requirements to be drawn up quickly which covered all disciplines involved, leaving more time for development.

Our general approach was to collect existing reports and similar documents that contain user requirements, to extract relevant information from them, and to present these needs in a coherent format. By requesting input from all partners within PARTHENOS, the project was able to cover a large number of past and current projects with potentially relevant reports, covering many subdisciplines from the fields of social sciences, humanities and cultural heritage. To make sure that use cases were practically applicable, the project decided to rely on real use cases as much as possible.

Consistency was ensured by applying a methodology as proposed by Alistair Cockburn in his book Writing Effective Use Cases.[12] According to this approach, a good use case description should have certain elements, such as: a descriptive statement of the goal; preconditions, describing what is necessary for the realisation of this goal; and the definition of successful and failed end conditions. This work resulted in sets of requirements, organised around the following five themes: data policies; standardisation; interoperability of data, services and tools; education and training; and communication needs. An example of a use case in the field of standards is: ‘Build a corpus of linguistic data for analysis’.[13] While this research scenario would be used primarily by linguists, the creation of a text corpus might be just as useful for a historian who works with a large body of text. This is where the added value of interdisciplinary cross-fertilisation becomes apparent. By facilitating the exchange of digital humanities research methods which are intrinsically widely applicable, humanities scholars are encouraged to embed methods of neighbouring disciplines into their own research.

Subsequently, this collection of use cases was handed over to the teams responsible for their implementation. Rather than taking the requirements and starting to implement them immediately, the Deming cycle (Plan-Do-Check-Act) was adhered to as follows.[14] By establishing an ongoing dialogue between the group of people that gathered the requirements and the implementation teams, the project continually verified whether initial user needs were correctly understood and whether the priorities as extracted from user documents were taken into account. Even during the implementation phase, the teams remained in close contact via meetings and workshops. Test versions and showcases were used to conduct quality assessments on the early versions of products. Lastly, assessments proved whether the final versions fulfilled the initial requirements.

The need for common policies and implementation strategies

One of the obstacles PARTHENOS has developed shared solutions for, is the scattered nature of policies concerning research data management in the humanities. Best practices and policies are often developed within traditional disciplinary boundaries. The reason for this is that research communities are most aware of their own disciplinary needs and of the distinctive character of the research assets they produce and use. An added benefit is that adoption is easier to encourage within rather than across disciplines. These are, on the one hand, the strengths of guidelines developed in close collaboration with a specific field. On the other, limiting policies and best practices to the confines of individual disciplines increases the risk of tunnel vision. Researchers who are active in the same field are more likely to interpret the information and the data they collect and produce in the same context, leaving much knowledge unexpressed or implicit. Also, either the perspective or the incentive to extend their insights beyond the limited focus of their research. It is when guidelines are not subject to scrutiny by researchers with a different focus, that they become prone to blind spots.

The Humanities disciplines are sometimes considered to be part of the 'long tail of science'. This means that data-centric research is perceived as less relevant for these fields than, for example, in physics or biology. The authors of this article consider this position to be only part of the story. While data are increasingly important in history, archaeology and related fields, the usual tools designed for big data research are often not appropriate for working with such datasets. The reason for this is that humanities scholars work with a large number of smaller datasets in many different formats and structures. This heterogeneity creates a stronger demand for tools that support both a better organisation as well as the transformation and re-expression of these data. Moreover, humanities research data does not yet have a well-established tradition of publication and is therefore often neither discoverable nor accessible. This creates a great risk to the sustainability of these data, which only increases when good practices in data management are not enacted. A lack of awareness around such practices in the humanities, lies at the heart of this risk of the creation of 'dark data'. The adjective 'dark' is a metaphor for such data being nearly unfindable. While 'big' humanities data do exist, their lack of findability can make them virtually non-existent for the researcher.[15]

As a cluster project, it has been PARTHENOS' aim to make optimal use of experiences gained throughout all disciplines. Shared best practices and formal policies make it possible to answer research questions in a more overarching way, enabling researchers from different universities and geographic areas to refer to the same policies in a shared language and mutual understanding. Also, researchers can more easily reuse each other’s results, allowing them to build on each other’s observations and knowledge. PARTHENOS focussed, therefore, the need for the integration of research data policies across the disciplines in the humanities, and therewith encourages interdisciplinary exchange. This is necessarily an on-going mission, for which one generic solution could never suffice. Nevertheless, different areas of high level compatibility could be identified through research over the current, distributed and previously unmapped landscape of research policies.

Three of the project outputs form an important beginning in providing solutions to this situation and can offer guidance in making research data more FAIR:[16]

1.) **The PARTHENOS Policy Wizard.** An interface which allows users to find information about policies which are relevant to their discipline and tasks. The user interface provides an intelligent categorisation that allows users to filter for policies that bear a relation to the researchers work.

2.) **The PARTHENOS Data Management Plan (DMP) template** was created, building on the Horizon2020 template while enriching and tailoring it with additional specifications from humanities disciplines. The content of this template was derived from a survey carried out among the consortium’s experts. It describes the life cycle of the data that is created, collected, archived and preserved by projects and RI’s in the humanities.

3.) PARTHENOS Guidelines. The PARTHENOS high-level guidelines are offered as common recommendations, aimed at building bridges between different, although tightly interrelated, fields and stakeholders within the Humanities. This encourages the harmonisation of policy definition and its implementation.

The need for information on standards

Like policies, standards constitute an important form of consensus on how Humanities research data can best be processed and stored. As a key element to interoperability and re-usability, they play a central role in the digital world at large. Contrary to policies, standards in themselves are non-legally binding methodological or technical specifications. Also, they adhere to the following three characteristics. Standards are:

1.) the result of a consensus building activity;

2.) publicly available, and;

3.) the object of a regular maintenance.

Conversely, through the mindset and practices standards create, they also build a common cultural background amongst the communities that have adopted them, increasing the probability and feasibility of further collaboration.

Apart from disciplinary standards, it is worth noting that there is also a need for generic (or horizontal) standards that provide a common background. Examples of this are the W3C XML recommendation; language codes (the ISO 639 series and the IETF BCP 47 document); the representation of time and dates (ISO 8601) and domain specific (or vertical) standards, such as the one developed in most ISO technical committees.[17]

The Humanities field is no stranger to the domain of standards. The Text Encoding Initiative guidelines have garnered significant support since their inception in various scholarly domains, ranging from epigraphy to literary studies and from history to lexicography. Additionally, ISO comprises a specific technical committee dedicated to language resources, which has provided various standards for the representation and annotation of linguistic content since 2002.[18] The natural interaction between scholars and cultural heritage institutions has also made standards such as EAD (Encoded Archival Description) part of the natural ecology of working with digital content in the humanities.[19] Similarly, the semantic representation of cultural heritage data has been standardised in CIDOC CRM (ISO 21127:2006), a high-level compatibility framework which is increasingly adopted in the humanities and e-sciences for data integration.[20]

Whereas some scholarly groups have been very active in defining and using standards in their practices, the project observed that – more generally – there is a lack of precise knowledge about standardisation, especially among newcomers to digital methods.[21] A solution devised to close this information gap is the Standardization Survival Kit (SSK), a tool which assists scholars in finding information on standards which are relevant to their specific scholarly activities. In line with the requirements gathered earlier in the project, research scenarios were drafted to help users on a step-by step basis to carry out a research scenario in a standardised manner.[22] The current scenarios illustrate both the importance of standards in the research process, as well as the usefulness of designing an online environment that allows researchers to access relevant reference material.[23]

The resulting digital environment was designed in close collaboration with digital specialists from the various communities represented in the PARTHENOS project. The current scenarios cover all types of scholarly domains and methodologies such as the management of field surveys and archival material, the creation of a textual or a musical corpus and the usage of laser techniques for conservation practice in heritage science.

The need for interoperability

A challenge that PARTHENOS has been in a unique position to tackle as a cluster project, was that of the integration of knowledge through the interoperability of data, generated within different institutes and RIs. To meet this aim, a conceptual model and data architecture have been developed to create and support trustworthy, sustainable, long-term integration processes.

Information management for RIs is both an epistemological and a technical challenge. The success of digital infrastructure largely depends on its fitness to the purpose of facilitating researchers in their collaborative development of knowledge.[24] One could argue that creating an RI involves building a community that as yet only partially exists towards a goal that as yet has not fully been understood. The information integration task is mishandled then if it is reduced to the question of which common system to adopt or what standard to enforce. Such ends can only be communally achieved and agreed upon, not presupposed or imposed.

Consequently, it was felt that a conceptual model was needed in order to support an information architecture that will not presuppose a community's identity and its direction, but to enable that community to arrange itself according to a well-formed picture of its commonalities and possibilities. Such an information system allows for the gradual identification of areas of common interest. Examples of integration activities are the analysis of resources that – when brought together – contain more information than the sum of their parts, and the creation of deeper integration of assets for areas where common methods and ideas are opportunate. This also entails a model that can provide means to track if data is reliable and well-maintained, allowing practical decisions to be made in terms of where to invest money and time.

PARTHENOS has therefore created a conceptual model of research infrastructure management itself, which models datasets, software, services, projects and actors and – most importantly – the contextual relations that exist between them. This conceptual model, the PARTHENOS Entities Model (PEM), provides the means to represent information on research assets in an accurate yet overarching way. In particular, it provides accurate modelling of the basic difference between kinds of services, such as hosting, curation and e-services. It also proposes a distinction between volatile and persistent digital objects, allowing the possibility to see the evolution of software and datasets.

These distinctions serve to facilitate more accurate and complex queries of the documented research assets. As a semantic framework, it does not impose a form of documentation, but provides a model which allows the translation of existing data about research assets into a common representation. Such a strategy keeps the barriers to adopting the model low. Meanwhile the model is aligned to CIDOC CRM in order to support interoperability with a wide variety of contemporary and future datasets. The architecture proposed alongside this model has, at its centre, a registry that documents what research assets exist, who holds them, how they are managed and where they can accessed. On top of the registry, it is envisioned that a number of deep aggregations can be built. These assets are continuously tracked through the registry and can serve as the motor for the development of new research across domains, building on the collective output of varying disciplines.

The cross-discipline interoperability described in the previous section allows researchers to have access to and to make use of the resources of other communities, thereby fostering true *interdisciplinarity*. But this is not enough: PARTHENOS has taken up the challenge of supporting *multidisciplinary* research, whereby members of different communities convene to the same ‘place’ to jointly pursue a common research question in shared projects. This is exactly what is envisioned in the development of the PARTHENOS Virtual Research Environment. It is commonly believed that such shared environments are key in enabling a new way of generating truly multidisciplinary knowledge. By making such a collaborative, underlying infrastructure available, built upon the cumulative results of projects in affiliated disciplines, PARTHENOS paves the way for new avenues of multidisciplinary research.

Skills, professional development and advancement

As described above, PARTHENOS did not only cover the technical aspects of RIs. The *human* network sustaining technical infrastructure and underlying data, as well as the act of making humanities researchers more aware of the potential of the computational methods that can be applied to analyse them, were considered just as important.[25] This is why a part of the project specifically focussed on offering the means to *learn about* both digital humanities research and the world of RIs.

Since the first ESFRI roadmap in 2006, there has not only been a rise in the coordinated development of research infrastructure in Europe. Significant changes have also been taking place elsewhere in the wider research ecosystem. Researcher careers have become, if anything, more precarious and less able to follow clear, pre-determined pathways. In the United States, this phenomenon has become known as the 'Alternate Academy,' or 'Alt-Ac',[26] and while it is widespread as a phenomenon, the discourse around it has largely risen out of the digital humanities, where interdisciplinary, applied and collaborative approaches open up wider perspectives than the ones which might be found in established disciplinary approaches.

The 'Alt-Ac' movement reflects North American conditions and prejudices, however, in its grass-roots organisation across a broad base of largely private institutions better able to differentiate in their approaches. In Europe, with its largely public research systems and regional linguistic variability, such a movement would be difficult to initiate. And yet researchers are still defining new and innovative pathways through their research and careers, and it is the research infrastructures, even more so than the universities, that are at the heart of this movement.[27] Accessing these opportunities requires a different perspective, different networks and different skills, however, than can normally be provided by higher education institutions. Digital research infrastructures are optimised for sustained work in teams, for creation rather than exploration, for the flexibility to harness technologies, policies and processes that are themselves still in development. This shift in requirement is evocative of how Rockwell and Sinclair describe the challenge of DH pedagogy, with research infrastructures representing a rethinking of teaching needs from the most fundamental level:

'One can think through a digital humanities curriculum in three ways. One can ask what should be the intellectual content of a program and parse it up into courses; one can imagine the skills taught in a program and ensure that they are covered; or one can ensure that the acculturation and professionalization that takes place in the learning community is relevant to the students'.[28]

Infrastructures are inherently committed to developing this third path, and the realisation of this model for the arts and humanities has been a key component of the PARTHENOS project. In particular, the project has approached this challenge through three mechanisms. The first of these is the PARTHENOS On-Line Training Suite, a collection of Open Educational Resources (OERs) developed by and for the infrastructure community.[29] Unlike other peer platforms, such as DARIAH Teach,[30] or the CLARIN VideoLectures portal,[31] the PARTHENOS Suite contains materials presented outside of formal curricula which reflect the core requirements and values within infrastructural work. This focus on the collaborative and integrative makes the materials in the Training Suite unique, and applicable to both independent learners and educators looking to expand their knowledge of such issues as what research infrastructures are, how they are managed, how to understand and manage humanities data, what collaboration means for the community, and other related issues. The organisation of PARTHENOS webinars has allowed for both additional learning opportunities, as well as the creation of newly created in-depth content for the Training Suite through their recordings.

On-line training alone is, of course, a blunt instrument, so PARTHENOS has also engaged with two other modes by which researchers hone their skills and build their networks in the digital age. The first of these is an analysis of Transnational Access**,** long a feature of European Research Infrastructures, but – in its original policy definition – not always a comfortable match with organisations that may focus on virtual access. Given the importance of problem-focussed, contextualised development of skills for the digital humanities, however,[32] it is critical that we understand how the virtual sits alongside the physical access, in particular for those researchers who may be more advanced in their careers, or at a turning point in their research project or approach. Secondly, PARTHENOS has partly closed the gap between formal education programmes, such as those in universities, and the knowledge creation and transfer modes of research infrastructures through a co-creation and exchange of curricular models with some of our university partners. Through these modes of engagement, the project aspired to create a more fluid transfer of skills and people between these two essential poles in arts and humanities research.

Communication, dissemination and outreach

Increasingly, individual researchers are encouraged, in many cases even obliged, to include a communication and dissemination plan in their research proposals. Both legal frameworks and moral appeals have inspired a significant increase in the creation of digital research data. The potential for research, which lies in this wealth of humanities data, is unequivocal. Scholars however, have indicated that the amount of pluriformity in the way these data are disseminated, increasingly creates 'disaggregated traditional scientific output' within an already fragmented communication system.[33]

In terms of dissemination, the challenge PARTHENOS faced was twofold. The first one was embedded in the design of the project itself, as the project has not only built, but also opened up an ecosystem, in which the tools and services described throughout this paper provide the envisioned benefits of integration and interoperability. By doing so, the project aspired to provide the means to approach this diversity in humanities research data and to deal with its tremendous breadth and pluriformity, thus providing an answer to the fragmented communication and dissemination systems. Secondly, PARTHENOS itself needed a deliberate dissemination strategy to allow for its own output to be discovered and used by scientific and professional communities. These two challenges existed by no means in isolation. The success of PARTHENOS' outreach activities determined to a large extent whether the project will have reduced 'complexity' – as Schroeder, Fry and de Beer describe the Humanities' fields diverse output – or added to the very problem.

A recurring concern around big projects is their 'one size fits all' approach. PARTHENOS, however, was very aware that both the information needs as well as the ways in which different stakeholders are best addressed are very different from one target group to the other. This is why, in the PARTHENOS' Communication Plan, different stakeholder groups were defined, allowing the project to match communication channels accordingly. However, these roles were never envisioned as a straightjacket. Rather, 'they merely exist to highlight the heterogeneous nature of the PARTHENOS' stakeholders, and to emphasise the need for a tailored approach to communication and dissemination, rather than act as a prescriptive classification'.[34]

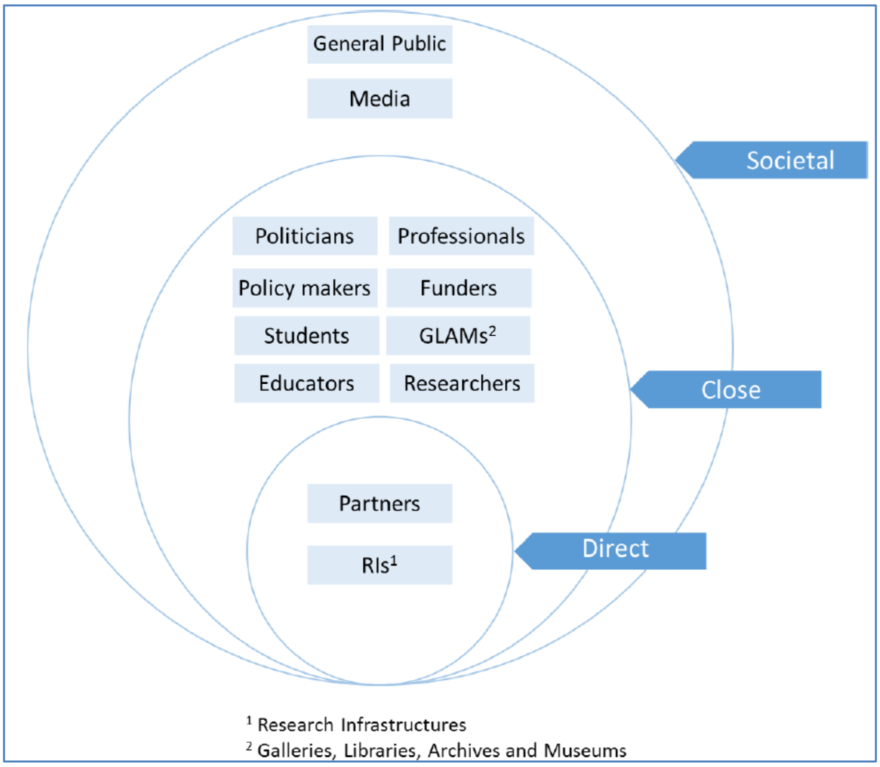
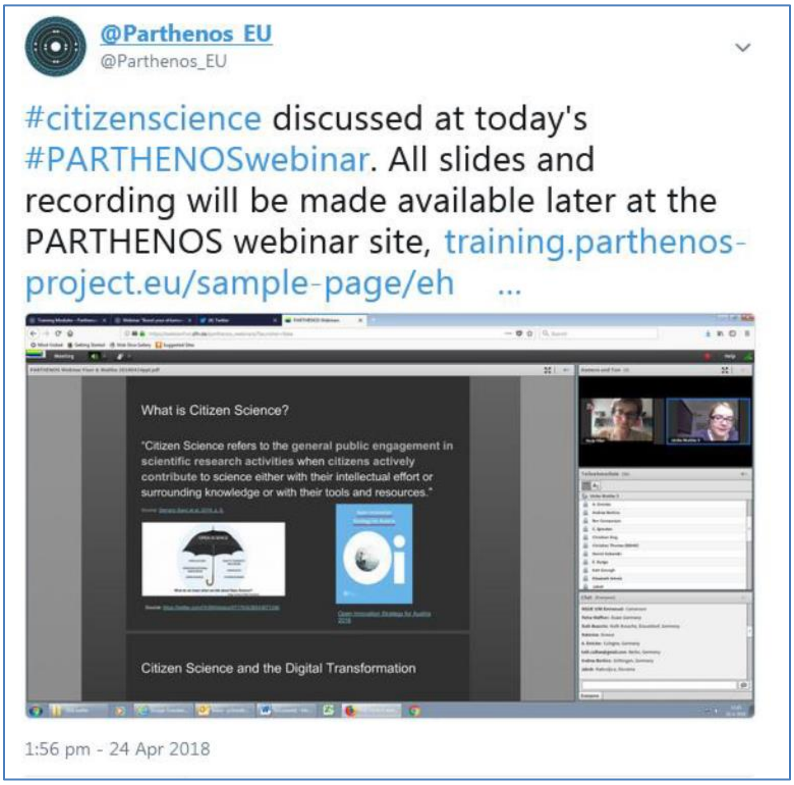


Figure 1: PARTHENOS stakeholder map

For a Humanities Research Infrastructure, the direct and the close audience formulated in the figure above could be regarded as evident. While researchers are its most prominent users, GLAM institutions are often important providers of data and expertise to humanities RIs. However, as an audience, society at large was considered just as important. Traditionally, interested individuals outside academia or museums have made very important contributions to (digital) humanities research, via their contributions to crowdsourcing events, hackathons or otherwise. PARTHENOS' dissemination channels, as well as its products and tools, were in no way restricted to academia. For that reason, PARTHENOS always considered it important that PARTHENOS' communication channels (the website, the newsletter, twitter etc.) are open to everyone. In that sense, PATHENOS never raised any institutional barriers and encouraged digital humanities citizen science – an umbrella term which includes all sorts of cooperation with interested individuals – just as much as research taking place in universities or research institutes.



@Parthenos_EU
@Parthenos_EU

#citizenscience discussed at today's #PARTHENOSwebinar. All slides and recording will be made available later at the PARTHENOS webinar site, training.parthenos-project.eu/sample-page/eh ...

What is Citizen Science?

"Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources."

Citizen Science and the Digital Transformation

1:56 pm - 24 Apr 2018

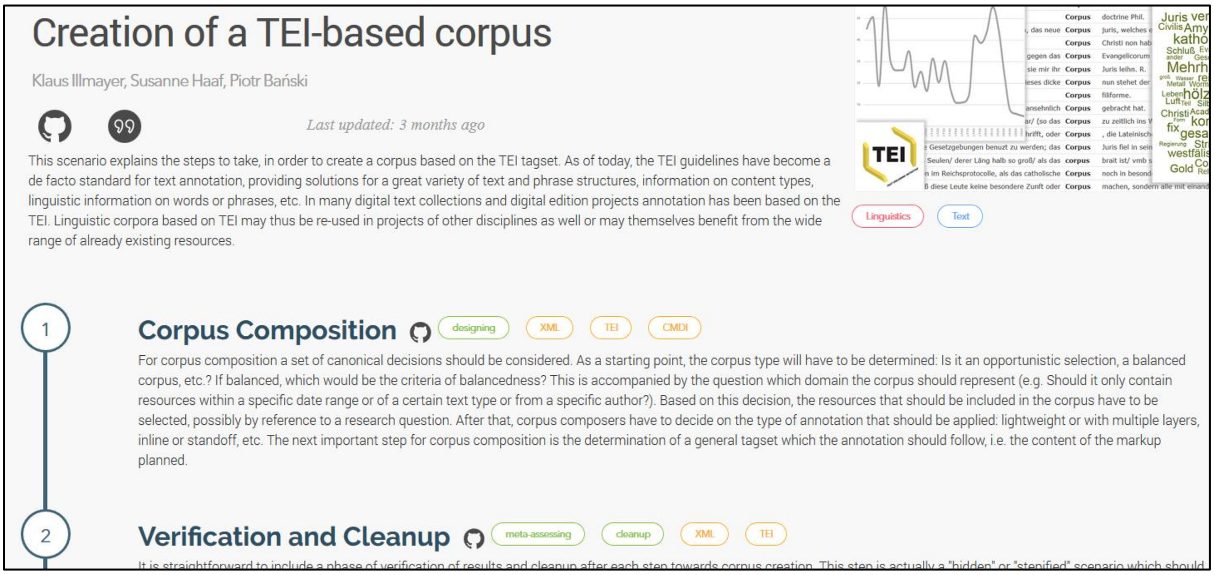
Figure 2: Citizen science, a topic discussed during PARTHENOS' webinar series

In conclusion, the open format of communication and dissemination PARTHENOS adhered to pays heed to three main points of criticism that large scale infrastructures have received in the past, namely that they can not successfully address a heterogeneous field, that they are an exclusive place and that, rather than promoting innovation are a restrictive force in that they are project-centred and enforce standards. However, through the different ways of communicating described above, PARTHENOS always aspired to be a 'loosely coupled ecosystem of services and activities', rather than a prescriptive force in an ivory tower.[35]

Conclusion: from requirement to application

If this paper intended to demonstrate anything, then it is that the similarity of the challenges that Humanities Research Infrastructures face are broad and wide-spread, but also comprehensible and – with the design of the right technical solutions – surmountable. By presenting a clear example, this conclusion will both demonstrate how products can support such a wide-spread, interdisciplinary field and how the requirements of such a diverse range of communities have found their way into the products we have developed over the past four years of the project.

Earlier in this article, 'Build a corpus of linguistic data for analysis' was provided as an example of a use case. This means that a community need in that specific area was identified, for which a shared solution was considered beneficial.



Creation of a TEI-based corpus

Klaus Illmayer, Susanne Haaf, Piotr Bański

Last updated: 3 months ago

This scenario explains the steps to take, in order to create a corpus based on the TEI tagset. As of today, the TEI guidelines have become a de facto standard for text annotation, providing solutions for a great variety of text and phrase structures, information on content types, linguistic information on words or phrases, etc. In many digital text collections and digital edition projects annotation has been based on the TEI. Linguistic corpora based on TEI may thus be re-used in projects of other disciplines as well or may themselves benefit from the wide range of already existing resources.

1 Corpus Composition

For corpus composition a set of canonical decisions should be considered. As a starting point, the corpus type will have to be determined: Is it an opportunistic selection, a balanced corpus, etc.? If balanced, which would be the criteria of balancedness? This is accompanied by the question which domain the corpus should represent (e.g. Should it only contain resources within a specific date range or of a certain text type or from a specific author?). Based on this decision, the resources that should be included in the corpus have to be selected, possibly by reference to a research question. After that, corpus composers have to decide on the type of annotation that should be applied: lightweight or with multiple layers, inline or standoff, etc. The next important step for corpus composition is the determination of a general tagset which the annotation should follow, i.e. the content of the markup planned.

2 Verification and Cleanup

It is straightforward to include a phase of verification of results and cleanup after each step towards corpus creation. This step is actually a "hidden" or "stepified" scenario which should

Figure 3: Creation of a TEI-based corpus as a research scenario in the SSK

The image above shows how this use case made its way from a requirement into a research scenario which can be consulted in one of the products described above, the Standardization Survival Kit. The screenshot presents the scenario ‘Creation of a TEI-based corpus’ which is organised in an easy to understand step-by-step format. This does not only allow researchers to not only learn how to build such a corpus and in what order. Also, along the different steps, additional resources are offered for further reading.

Therewith, this article does not only attempt to demonstrate how tools can support the digital humanities field at large. Lastly, it has set out to provide insight in how projects can develop tools for researchers while keeping the needs of user communities in mind. The constant dialogue between two teams – with one focussing on user requirements and the other on their implementation in the products that were under development – has proven to provide an organised and controlled mode of delivery. For that reason, we would recommend any project to define clear criteria and to verify constantly whether the products that are being build live up to the desired functionality.

[1] European Commission, *European Research Area (ERA)*, https://ec.europa.eu/info/research-and-innovation/strategy/era_en, last accessed 15 March 2019.

[2] ESFRI, *European Roadmap for Research Infrastructures – Report 2006*, https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf, last accessed 15 Mar 2019.

[3] W*ebsite DARIAH*, https://www.dariah.eu/, last accessed 20 Sept 2018.

[4] *Website CLARIN-ERIC*, https://www.clarin.eu/, last accessed 20 Sept 2018.

[5] *Website CENDARI*, http://www.cendari.eu/, last accessed 20 Sept 2018.

[6] *Website EHRI*, https://www.ehri-project.eu/, last accessed 20 Sept 2018.

[7] *Website ARIADNE*, http://www.ariadne-infrastructure.eu/, last accessed 20 Sept 2018.

[8] *Website IPERION-CH*, http://www.iperionch.eu/, last accessed 20 Sept 2018.

[9] C. Leathem 'Survey and Analysis of Humanities and Social Science Research at the Science Academies and Related Research Institutes of Europe', in A. Duşa, D. Nelle, G. Stock and G. Wagner, eds., *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences* (Berlin, 2014), 39–43. Cited here at 42.

[10] Terminology borrowed from: G. Lauer 'Challenges for the Humanities: Digital infrastructures', in Duşa, Nelle, Stock and Wagner, eds., *Facing the Future*, 35–38. Cited here at 36.

[11] *Website PARTHENOS*, http://www.parthenos-project.eu/, last accessed 20 Sept 2018.

[12] A. Cockburn, *Writing Effective Use Cases* (1999) https://www.infor.uva.es/~mlaguna/is1/materiales/BookDraft1.pdf, last accessed 15 Mar 2019.

[13] S. Drude et al., *D2.1 Report on User Requirements*, http://www.parthenos-project.eu/Download/Deliverables/D2.1_User-requirements-report-v2.pdf, last accessed 15 Mar 2019, 117-118.

[14] M. Pietrzak and J. Paliszkiewicz, 'Framework of Strategic Learning: The PDCA Cycle', *Management*, 2 (2015), 149-161.

[15] P.B. Heidorn, 'Shedding Light on the Dark Data in the Long Tail of Science', *Library Trends*, no. 2 (2015), 280-299.

[16] The FAIR principles provide an important frame of reference for the output PARTHENOS generates on the topic of research data management. See: M.D. Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, no. 3 (2016).

[17] *Website International Organisation for Standardisation (ISO) Technical Committees*, https://www.iso.org/technical-committees.html, last accessed 20 Sept 2018.

[18] *Website ISO standards catalogue on language resource management*, https://www.iso.org/committee/297592/x/catalogue/p/1/u/1/w/0/d/0, last accessed 20 Sept 2018.

[19] *EAD project website*, https://www.loc.gov/ead/, last accessed 20 Sept 2018.

[20] G. Bruseker, N. Carboni, and A. Guillem, 'Cultural heritage data management: the role of formal ontology and CIDOC CRM', in: M. Vincent, V.M. López-Menchero Bendicho, M. Ioanniders and T.E. Levy eds., *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data* (Cham, 2017), 93-131.

[21] K. Illmayer and M. Puren. 'How to work together successfully with e-Humanities and e-Heritage Research Infrastructures' *PARTHENOS eHumanities and eHeritage Webinar Series*, http://training.parthenos-project.eu/sample-page/ehumanities-eheritage-webinar-series/webinar-work-with-research-infrastructures/, last accessed 20 Sept 2018.

[22] This is an example of the interaction between the team which collected requirements and the team which builds a tool to fulfill the addressed needs (see under 'an approach based on the needs of communities').

[23] For additional background, please see: L. Romary, et al., *D4.1 Standardization Survival Kit (Draft)*, https://doi.org/10.5281/zenodo.2668403, last accessed on 6 May 2019; L. Romary, et al., *D4.2 Report on Standardization (draft)*, https://doi.org/10.5281/zenodo.2668414, last accessed on 6 May 2019.

[24] This is for instance illustrated by the juxtaposition of 'interoperability' (a technical accomplishment) and 'interchange' ('true' human-mediated interoperability by understanding). See: Martin Holmes, 'Whatever happened to interchange?', *Digital Scholarship in the Humanities* no. 1 (2017), 63-68, https://doi.org/10.1093/llc/fqw048, last accessed 6 May 2019.

[25] The vocabulary of a technical infrastructure and a human network borrows heavily from EHRI's mission. *European Holocaust Research Infrastructure – Mission Statement*, https://www.ehri-project.eu/about-ehri, last accessed 23 July 2019.

[26] B. Nowiskie (ed)., *#Alt-Academy: 01. Alternative Academic Careers for Humanities Scholars* (2011) http://mediacommons.futureofthebook.org/alt-ac/sites/mediacommons.futureofthebook.org.alt-ac/files/alt-academy01.pdf, last accessed 29 Mar 2019.

[27] J. Edmond (2019) 'Cultures of Scholarship and Work: Reflections on Infrastructure as the European alt-ac' In: M. Gold and L. Klein, eds., *Debates in the Digital Humanities 2019* (Forthcoming).

[28] G. Rockwell and S. Sinclair, ‘Challenges for the Humanities: Digital infrastructures’, in B. Hirsch, ed., *Digital Humanities Pedagogy: Practices, Principles and Politics* (Cambridge, 2012), 177–212. Cited here at 178.

[29] *Website PARTHENOS On-Line Training Suite*, http://training.parthenos-project.eu/, last accessed 29 Mar 2019.

[30] *Platform DARIAH Teach*, https://www.dariah.eu/teach/, last accessed 29 Mar 2019.

[31] *CLARIN VideoLectures Portal*, http://videolectures.net/clarin/?q=clarin, last accessed 29 Mar 2019.

[32] S. Antonjević, *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production* (Basingstoke 2015).

[33] R. Schroeder, J. Fry and J.A. deBeer, ‘e-Research Infrastructures and Scientific Communication’, *Proceedings of the IATUL Conferences*, paper 28.

[34] R. Speck, S. Sbarbati, S. Bassett, F. Niccolucci, P. Drenth, D8.2 Initial Communication Plan. 31 July 2015. http://www.parthenos-project.eu/Download/Deliverables/D8.2_Initial_Communication_Plan.pdf, last accessed 12 April 2018, p.15.

[35] This concept is derived from: T. Blanke, C. Kristel and L. Romary, ‘Crowds for clouds: recent trends in humanities research infrastructures’ in: A. Benardou, E. Champion, C. Dallas and L. Hughes, eds, *Cultural Heritage Infrastructures in Digital Humanities* (Abingdon, 2017) 48-62.