TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship

Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Kuster, Malcolm Illingworth

Journal of the Text Encoding Initiative

Issue 5 (June 2013)
TEI Infrastructures

Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Küster and Malcolm Illingworth

TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship

Electronic reference
Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Küster and Malcolm Illingworth, « TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 25 June 2013, connection on 10 December 2013. URL : http://jtei.revues.org/774 ; DOI : 10.4000/jtei.774

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/774
Document automatically generated on 10 December 2013.

Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Küster and Malcolm Illingworth

TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship

1. Introduction

1 In recent years, a variety of initiatives have been funded with the aim of producing software tools or environments of a type variously known as *virtual research environments*, *research infrastructures*, or *cyberinfrastructures*. These initiatives vary in their scale, specialization, scope, and level of funding. One issue that they face in common, however, is that of sustainability: how can the continued—and useful—existence of a system or tool be guaranteed, or at least facilitated, once a project's funding has been spent? In this paper, we examine how such sustainability has been enabled, in the particular case of infrastructures for textual scholarship, in the context of three international projects: TextGrid,[1] TEXTvre,[2] and DARIAH[3]. Firstly, we will address the inter-project collaboration and cross-fertilization between TextGrid and TEXTvre, including architectural decisions and shared data infrastructures, and investigate how the projects benefited from the exchange. We will then discuss how this existing collaboration can be taken forward by the loosely-coupled and distributed framework being developed by the DARIAH community, and how it can serve as a model for the sort of collaborations that DARIAH plans to enable.

2 TextGrid is a longstanding German initiative to establish a virtual research environment for several text-based humanities disciplines. It has been running for many years and its current aims are to achieve a high level of sustainability of its work and to collaborate with international partners to enhance its offerings. The Digilib[4] project is one example of a scholarly editing environment within TextGrid. Digilib is noteworthy for its strong focus on an existing community, and it has led to various attempts within other communities to adopt and replicate its methodologies and technologies.

3 One of TextGrid's international collaborations has been with the UK-based and JISC-funded TEXTvre project, which aimed to adopt and institutionalize TextGrid's solutions for the well-established community for digital textual scholarship in the Department of Digital Humanities at King's College London. To this end, TEXTvre spoke to King's researchers engaged in these activities, analyzed their responses, and enhanced the TextGrid services to suit their needs, for example by integrating better support for automated text analysis.

4 The Digital Research Infrastructure in the Arts and Humanities (DARIAH) is building a virtual bridge between different humanities and arts resources across Europe. Initially funded under the ESFRI programme,[5] DARIAH combines various national infrastructures—from the Dutch national support infrastructures for e-humanities to the German e-humanities infrastructure TextGrid. It also aims to help other EU countries establish their own arts and humanities e-infrastructures, and to achieve new modes of collaboration between computing and humanities based on existing communities of practice such as the one that has emerged around TextGrid. However, there is no generic way to use a research infrastructure; it needs to be oriented and customized towards the needs of emerging communities. Services in such an infrastructure are useful only if they are built around such communities of practice.

5 This paper is structured as follows. In section 2 we describe the TextGrid and TEXTvre projects and examine how they relate to each other. Section 3 discusses their collaboration on tools and services, and section 4 their deeper collaboration on data infrastructure, before we examine in section 5 how these could be further supported within the context of DARIAH. Finally, in section 6 we address ongoing collaborations on sustainability and on tools and services.

2. Background: TextGrid and TEXTvre

6 The TextGrid research group, a consortium of ten research institutions in Germany, has developed a virtual research environment (VRE) for researchers in the arts and humanities. It provides services and tools for the digital analysis of text data and supports the curation of research data using grid technologies. Research partners—including libraries and data centers as well as universities and research institutions—are collaborating in a community-driven process that is funded by the German Ministry for Education and Research (BMBF). As part of the German Grid Initiative D-Grid,[6] TextGrid maintains a common Grid Resource Centre in Göttingen (GoeGrid)[7] together with grid projects in physics, medicine, astronomy, and climate research.

7 An essential aspect of the TextGrid infrastructure is that all its technical developments have been accompanied by the definition of precise recommendations, most of them based on the Text Encoding Initiative (TEI) Guidelines. These Guidelines, which for example establish precise constraints on stand-off annotation mechanisms or on the basic representation of dictionary entries, have contributed to the propagation of the idea of an infrastructure as comprising both technical and intellectual services. In particular, it provides a range of tools for supporting the production of digital editions based on TEI.

8 Currently, TextGrid consists not just of its original communities—textual philology and linguistics—but also many others, for example art history, classical philology, and musicology, all of which have extensive interests in textual editing and annotation. TextGrid was conceptualized as an infrastructure that enabled new communities, together with their tools, to be easily integrated and also to take part in the development process. TextGrid has tried to put these communities first, and thus has only gradually and carefully introduced changes to the technologies and methodologies that it uses in its VRE.

9 As detailed by Neuroth, Lohmeier, and Smith (2011), the TextGrid VRE consists of two main components: the TextGrid Lab(oratory), which serves as the entry point for users, and the TextGrid Rep(ository), which provides access to a shared data archive. The TextGridLab provides common functionalities in a sustainable environment in order to increase reuse of data, tools and services, and the TextGridRep enables researchers to publish and share their data in a way that supports long-term availability, interoperability, and reusability.

10 TEXTvre built on the success of TextGrid and worked towards an institutionally-focused VRE that addressed the requirements of a wider community of practice of textual scholars and could be integrated with institutional infrastructure. Its approach to this involved engagement with researchers based in and collaborating with staff in the Department of Digital Humanities at King's, exploring their research practices and integrating user tools to support and enhance these practices.

11 The Department of Digital Humanities (DDH; formerly known as the Centre for Computing in the Humanities) is an academic department in the School of Arts and Humanities of King's College London and the home of many multidisciplinary research projects in the digital humanities (including many text-based ones) carried out in collaboration with a range of humanities departments, both at King's and elsewhere. TEXTvre's engagement with humanities researchers was mediated through this department and focused on research projects in which DDH was a collaborator.

12 One of the main aims of the TEXTvre project was the analysis of existing research practices to identify processes that are typical of an institutional environment for digital humanities. It was expected that there would be significant differences at DDH in comparison with a national infrastructure such as TextGrid. Three case studies, which were considered to be typical of the critical edition projects carried out within the department, were selected:

1. Early English Laws[8] produced new editions and translations, both online and in print, of all English legal codes, edicts, and treatises produced till the Magna Carta in 1215. The project operated in an extremely heterogeneous document environment, and the researchers were distributed across the UK. It also had very specific descriptive,

administrative, and technical metadata requirements that weren't then accommodated in TextGrid.

2. The Gascon Rolls Project[9] addressed texts that form a fundamental primary source for the study of the administrative, political, and economic history of the French region of Aquitaine under English rule. The source documents are somewhat inaccessible in the UK's National Archive, so the project developed an online edition with summary translation, high-quality images, and detailed indexes. It worked within a well-developed tradition of editing historical records (at least printed editions) but required the creation and support of an ontological layer.
3. The Inscriptions of Roman Tripolitania[10] produced an online edition of the inscriptions, mainly in Latin, from the Roman province of Tripolitania (in modern Libya). There is an established print-based tradition for the editing and publication of inscriptions, and a reasonably well-established extension of this to the digital domain, EpiDoc,[11] has grown out of this tradition, in part as the result of work by the IRT team.

13 Based on an analysis of these projects, TEXTvre identified the kinds of services and tools that could be reused from the TextGrid environment, as well the gaps in provision.

3. Working Together on Tools

14 As discussed above, TextGrid is built on a distributed architecture and offers its users access to an ecosystem of different services and content; these tools will be the focus of this section. TextGridLab, TextGrid's Eclipse-based user interface, integrates a set of interactive tools such as the XML Editor, the Text-Image Link Editor (German "Text-Bild-Link-Editor", TBLE) and the Text-Text Link Editor (discussed in section 6). The XML Editor is one of the core functionalities of TextGridLab, allowing users to switch easily between a more technical view with tags and attributes and a structural view that is oriented towards standard text-editing applications. The Unicode Character Table enables the user to search, copy, and insert symbols from the Unicode character set. The Text-Image Link Editor supports the XML Editor by linking text sequences with image sections in order to create files that contain text elements and topographic descriptions.

15 TextGrid offers various utility tools that support the XML annotation process. The Dictionary Search Tool, for instance, provides an integrated search interface to a number of different dictionaries within the TextGrid VRE. The Wörterbuchnetz dictionary network[12] at the University of Trier has been integrated into the interface for this purpose. The CollateX[13] tool is another utility tool for comparing two or more files encoded in XML (including TEI) and annotating any differences.

16 The Text Publisher Web tool, on the other hand, can be used to present project results and publications on a project website. The tool is part of a TextGrid toolkit for web publishing. For more specialized publishing tasks, a module called XML Print[14] is integrated into TextGrid to allow printing of scholarly texts with complex typesetting layout requirements based on XML data, with a particular view towards the needs of critical editions and dictionaries. Finally, the Bibliography Tool can import bibliographical data from existing data inventories and be used to capture, process, and administer bibliographies. Users can also export bibliographical data into certain standard formats (for example, TEI and MODS).

17 With the addition of new user communities, TextGrid has added new tools and customized existing ones. The Note Editor illustrates MEI[15]-encoded scores and displays them in a simplified format. Unique features of the Note Editor, such as the visualization of variants, are located in the editorial field. Another new tool is the Gloss Editor, which facilitates the description of specific text structures in gloss comments, the synoptical presentation of different gloss comments regarding the same text, and the synoptical presentation of gloss comments and related digital manuscripts.

18 TEXTvre wanted to go beyond these tools and services and interface with tools relevant to the local environment. These included alternatives to standard TextGrid tools, such as the commercial XML editor <oXygen/>, and services for linguistic analysis based on the open

source framework GATE (Cunningham, Maynard, and Bontcheva 2011; Cunningham et al. 2002).

19 <oXygen/> is the XML editor of preference for DDH projects, so the TEXTvre team considered accommodating this to be an important candidate for TEXTvre. The <oXygen/> Eclipse plugin, which supports all of <oXygen/>'s basic XML editing functionality,[16] was integrated into TextGridLab, allowing users to create TEI files with <oXygen/> instead of the TextGrid editor.

20 TEXTvre also integrated completely new tools and services. The Solr[17] plugin, for instance, allows a user to index one or more documents that exist in the repository, search over the index, define a Solr schema for specific searching requirements, and browse using facets. There was some concern that these features duplicate existing TextGrid search functionality and may lead to confusion; however, it was felt useful to have a client-side, customizable faceted browsing tool for project-specific search.

21 Some work was carried out on integrating GATE services with the TextGridLab client platform to provide more advanced text analysis functionalities than those available in TextGrid. GATE is an information extraction and semantic annotation framework developed at the University of Sheffield, and this aspect of the work was intended as a demonstrator of the potential for integrating such services. Specifically, TEXTvre implemented named-entity recognition (NER) web services for German and English. These services take as their input an XML document (typically TEI XML) and return the same document with additional XML tagging for the named entities. The services contain slightly modified versions of GATE's German NER pipeline and the ANNIE information extraction components.[18]

22 In addition to creating tools and services, TEXTvre and TextGrid have also engaged in a fundamental collaboration on the underlying data infrastructure of both projects, which we describe next.

4. Data Infrastructure Collaborations

23 The second main component of TextGrid is the TextGridRep, which provides access to a repository infrastructure for long-term preservation of data based on grid technology. Researchers can decide how and with whom their data will be shared by using sophisticated rights management services. Findings and research data can be published directly from the TextGridLab to the TextGridRep via a publishing process that guides researchers in preparing the data for long-term accessibility.

24 The TextGridRep middleware consists of various components for handling files in the data grid, for rights management, for metadata management in an XML database, and for relations in an RDF triple store. TextGrid is not a closed system but rather an open platform that enables scholars to adapt the environment to their needs. Owing to its modular structure (Küster et al. 2007), it is easy to integrate external web services (such as the dictionary network "Wörterbuchnetz" referenced above), and the layered architecture also allows access to the services via graphical user interfaces other than TextGridLab.

25 It did not meet the requirements of the researchers at King's College London simply to reuse the German Grid infrastructure. King's had already started to develop its own repository infrastructure based on the Fedora repository technologies (Lagoze et al. 2006), and it was decided to integrate the VRE with this instead. Integrating TextGrid with Fedora essentially meant implementing management of digital objects in a Fedora object repository instead of using the German D-Grid (or similar grid-based resources). The TEXTvre team wanted to use Fedora to hold complete representations of the objects—not just to use it as a data store—so as to be able to use Fedora as the basis for various modes of publication and reuse of the content. This also requires storing inter-object relationships in Fedora, thus implementing the TextGrid object model by means of the Fedora Digital Object Model and the underlying Content Model Architecture (Lagoze et al. 2006). Within TextGrid, object management is carried out by the TG-CRUD service,[19] which is modular and has a defined storage interface for implementing access to an object repository. For the TEXTvre implementation, the team developed a new component implementing this interface and allowing objects to be managed in Fedora instead

of D-Grid, making use of the Fedora REST API instead of writing to D-Grid using JavaGAT,[20] the Java-based grid interface. This in itself was relatively straightforward, and is illustrated in figure 1.

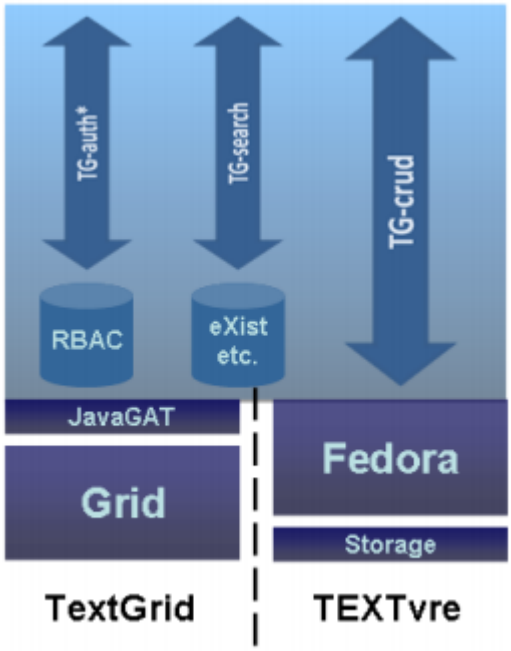

Figure 1. TextGrid and TEXTvre data infrastructure

26 However, replicating TextGrid relationships in Fedora proved to be difficult because of the various ways in which relationships and object versioning are managed within TextGrid, and so a simple mapping of the RDF in the TextGrid triple store to the RDF in Fedora did not work. Specifically, in TextGrid, the relationships are stored in a Sesame RDF database,[21] and the objects themselves are stored elsewhere, whereas in Fedora, relationships are stored in a reserved datastream (called RELS-EXT) of the source object, where the target object must already exist within Fedora.

27 An additional complication is that edition projects as a whole are not actually modelled as objects in TextGrid (i.e., they are not managed by TG-CRUD); instead, they are LDAP constructs referenced in TextGrid relationships. As there are no objects in TextGrid corresponding to projects, an attempt to store the relationship between an object and a project in Fedora fails as the target object does not exist.

28 Other issues arose from TextGrid's approach to versioning. In TextGrid, different versions of an object can be created, and each version is maintained as an independent object and can be edited and changed at any time; however, no change management is available on the individual objects. TextGrid maintains the concept of a current version of an object, and the "current" version number is kept as a field in the object's metadata in the eXist database. For example, an initial version of an object might have an identifier textgrid:45679.0, indicating version 0. Creating a new version would result in a new object with identifier textgrid:45679.1, and so on. In relationships between objects, TextGrid can involve both specific versions of an object (e.g. textgrid:45679.2) and "unversioned" objects (e.g. textgrid:45679), where the most recent version of an object is assumed. Fedora, on the other hand, has an entirely incompatible concept of versioning, in which object versions are implemented using a timestamp that captures the state of the object at a particular point in time. Fedora's own versioning mechanism therefore cannot be used to implement TextGrid versioning. To address these issues, a subclass of the TextGrid component providing relationship storage was implemented for TEXTvre, with the following processing:

- Relationships are received as an input stream of triples in N3 format,[22] which are processed into individual triples using the Jena library.[23]

- As projects do not exist as objects in TextGrid, an explicit check is made for "inProject" relationships, and a dummy project object is created in Fedora (if it does not already exist).

29 Each time that a relationship is added to Fedora, the component checks that both of the objects in the relationship already exist in Fedora. If either or both of the objects do not exist, placeholder objects are created and the relationship is stored referencing these. A user searching via Fedora can easily see whether an object referenced in a relationship is versioned or is an unversioned placeholder. This part of the project led to additional funding from the Software Sustainability Institute (SSI),[24] an EPSRC-funded[25] body that helps to increase the sustainability of open-source research software. This allowed TEXTvre to streamline the process of installing and configuring an instance of the software, develop a virtual machine image that allows a TextVRE installation to be run "out of the box", and increase modularization of the source code. The work was carried out in collaboration with EPCC[26] at the University of Edinburgh, as well as with the TextGrid team at Göttingen and DAASI.[27]

5. Sustainability: The DARIAH Initiative

30 As is evident from these discussions, large research infrastructures such as TextGrid do not consist of a single solution and a single toolkit of services, but are essentially a combination of smaller, often standalone, services that a user can combine at will. This is particularly the case in digital environments for the arts and humanities. DARIAH recognizes this, and thus a key feature of its provisions is a virtual "social marketplace", a framework that offers projects such as TextGrid and TEXTvre a forum for exchanging such tools and services, and that supports advanced collaboration across diverse networks and specialized service providers. As detailed by Blanke et al. (2011), such a marketplace has three main pillars:

1. Open APIs exposing reusable services
2. Composition and aggregation facilities allowing researchers to build on these services
3. Promotion of applications that are based on these services, for a variety of use cases

31 The rules and procedures implemented by DARIAH for its marketplace determine how and where participants can make their tools and services available. It is an open environment where services can be exchanged, but of course in any such environment trust is a major issue, and some kind of compliance framework is required to ensure the quality of the exposed services. Within the DARIAH community, a service can simply be documented without certification, in which case users cannot assume that it is reliable, or else it can be fully certified as compliant by the DARIAH organization (Blanke et al. 2011). Compliance with DARIAH guidelines ensures that data and services can be contributed by decentralized parties while at the same time ensuring that they are both accessible and trustable across the DARIAH ecosystem. Thus, while DARIAH is not concerned intrinsically with TEI, it provides a framework for making available and sustaining the TEI-related functionality provided by other initiatives such as TextGrid.

32 In order to scale tools and services from projects such as TEXTvre and TextGrid to a sustainable European collaboration, DARIAH has developed a flexible organizational model that enables pan-European collaboration. DARIAH has set up a number of Virtual Competency Centres (VCCs), each of which is based on topics that, based on deep experience with digital arts and humanities projects, are part of the challenges any such project faces:

- an e-Infrastructure VCC, which is concerned with the technical platform
- a VCC for liaising with research and education communities, thus ensuring that research activities are directly embedded in the infrastructure
- a VCC for scholarly data, which supports the long-term availability of data
- a VCC that works on the impact of the tools and resources that DARIAH incorporates

33 DARIAH provides these VCCs as platforms for collaboration. The DARIAH partners between them have extensive experience with meeting the kind of challenges that both TextGrid and TEXTvre have faced; however, the flow of expertise is not one-way, and DARIAH can learn

from the collaboration experiences of these projects. Customizing one's own storage solutions is often a very involved task that needs careful planning and organization. DARIAH anticipates that many digital humanities research groups and centers will have to comply with their own institutional requirements, while for many countries there will be other dedicated national services for supporting the work of data publishing and preservation. Therefore, a common DARIAH data layer will not consist of a single data infrastructure but rather will comprise a set of services, either for reusing one of several existing solutions or for providing a new solution, and thus will build on the expertise of the DARIAH partners. The experience of the collaboration between TEXTvre and TextGrid can be seen as exemplary in this context.

34 While variety is to be expected at this level, the diversity of local and national data services will be brought together into a common European data layer through a toolkit of lightweight services that allows, for example, cross-referencing and discovery across various data repositories. The existing DARIAH Persistent Identification (PID) service is a good example of such a lightweight solution (Blanke et al. 2011). DARIAH recognized early on that the use of PIDs within the new infrastructure is imperative; when a researcher cites an article or dataset, whether digitally or in hardcopy, they need to be assured that the citation itself will always lead to the original resource cited. The creation of relationships between resources, for example between a researcher and the articles they have published, requires a similarly persistent mechanism. Many DARIAH members already provide their own PID solutions; this is the case for TextGrid and TEXTvre, for example. The DARIAH PID service has been deliberately designed to accommodate these local solutions, scaling them to a European level, and allowing DARIAH PIDs to be used in the heterogeneous environment of existing archives and VREs, for example in the Fedora-based repository used by TEXTvre.

6. TextGrid–TEXTvre Collaboration: Next Steps

6.1 Sustainability

35 TextGrid started its third funding period, which will last for three additional years, in June 2012, and the main focus of this period is on sustainability, specifically how to ensure that TextGrid will be able to provide services and tools after the project has come to an end. The second major concern of this phase is to address the scalability and performance issues of the TextGrid middleware: increasing numbers of research groups are joining TextGrid, leading in turn to increasing numbers of concurrent users, so a stable, sustainable, and trusted infrastructure has therefore become paramount.

36 TextGrid's sustainability issues are not just technical; a key aspect is the financial and organizational sustainability of running the system after funding runs out in 2015. While "big science" disciplines, such as astronomy, climate research, or particle physics, have well-developed approaches to setting up the financial and political structures required for such infrastructures, for example a legal entity with a reliable budget, until now there has been no such model to follow in the arts and humanities communities. The roadmap developed for the sustainability of TextGrid, which has been produced at the request of the German Ministry of Research and Education, will serve as a blueprint for the sustainability of other humanities research infrastructures in Germany. Moreover, TextGrid is working closely with DARIAH to determine the extent to which DARIAH services, such as the PID service, the authorization/authentication/identification (AAI) infrastructure, and storage solutions for data curation can be leveraged to support this sustainability.

37 TextGrid is likely to be sustained through a hybrid financial model, with some of the budget provided by the central government, supplemented by funding from German state governments that have already set up their own digital humanities initiatives. Other stakeholders could also provide support to VREs such as TextGrid by providing central services in kind, while researchers who plan to use TextGrid could request additional money in their funding proposals. The main federal funding agencies in Germany, the BMBF (Bundesministerium für Bildung und Forschung) and the DFG (Deutsche Forschungsgemeinschaft), have agreed to support this idea, which has operated in the sciences for many years. TextGrid is thus entering its operational phase. After the first phase, which focused on conceptual issues, and

the second phase, which saw most of the development, TextGrid is now targeting its own long-term future, focusing on sustainability, trust, and financial viability. This future also includes education and training, such as the development of professional teaching materials, special lectures within BA and MA curricula, and expert workshops on innovative topics in collaboration with national and international research and developer communities.

6.2 New Services: Embedding Intertextual Links

38 Extensibility in the face of new requirements is an important attribute of any infrastructure, and both TextGrid and TEXTvre have given rise to follow-on projects that addressed specific research questions. One such set of requirements for additional functionality arose from the needs of the SAWS (Sharing Ancient WisdomS) project[28] (Jordanous et al. 2012), which addressed the digital publication of medieval wisdom literatures in Greek and Arabic. These texts, also termed *gnomologia*, are anthologies of "wise sayings", encapsulating moral and social guidance or philosophical ideas. Gnomologia also formed a key route via which ideas were disseminated over a large area, and across different cultures, often over the course of many centuries.

39 While the publication of such collections using TEI was one objective of the SAWS project, a distinguishing characteristic of these texts—which both makes them of particular interest to scholars and raises particular challenges for existing models of digital publication—is that the sayings or excerpts from which they are constituted are borrowed, modified, and recombined in different ways as new manuscripts are written on the basis of older ones. A key requirement of a scholarly editing and publication infrastructure for such texts is that it should support identification, representation, and exploration of the relationships that occur between excerpts, both within texts and across multiple texts. For example, one anthology may repurpose excerpts from a number of existing anthologies, changing them in various ways; sayings may be translated, and often changed in subtle ways in the process; and many of them can be traced back to known "source texts", often classical texts or the Bible.

40 Consequently, as well as producing digital editions of the manuscripts, SAWS aimed to build up a corpus of links between textual excerpts. As part of the markup process, scholars need to identify excerpts and relationships and mark them up within the TEI document. Traditionally such information would form part of the human-readable *apparatus criticus*; instead, SAWS is producing a network of interrelated excerpts formalized using an RDF-based model, in addition to the more standard TEI approach. As such relationships between texts and text fragments become more important to the work of scholars, tools are needed for editing and managing them in a persistent manner (Hedges et al. 2012).

41 These requirements are being addressed by the Text-Text Link Editor (TTLE), a tool developed at the TextGrid partner Fachhochschule Worms that handles links between arbitrary fragments of XML documents. TTLE can be integrated as a plugin into TextGridLab, thus providing textual scholars with a consistent user experience and workflow for both standard TEI editing and working with intertextual links. TTLE's persistence layer uses the linking functionality already provided by TEI; although there are other standards available to describe text linking—for example, the Open Annotation Core Data Model[29]—using TEI allows TTLE to make use of the services already available in the TextGrid repository.

42 The mechanism TTLE uses to mark up fragments depends on the status of the document. For writeable documents, it simply inserts specific tags, whereas for write-protected documents it either uses character offsets or allows an identifier to be assigned to any text enclosed in an XML tag. In either case, the information is stored in TEI format, either in the document itself or in a separate file. The current version allows users to add comments to a link, and subsequent releases will offer a way of specifying link types for easy sorting, searching, and filtering.

7. Conclusions

43 This paper has presented the collaborations between the TextGrid and TEXTvre projects in Germany and the UK, and examined how DARIAH might operate to facilitate their collaborations. We have shown how TextGrid emerged from the specific needs of a German

community and was then adopted and enhanced for the needs of the scholarly community at King's College London. These two virtual research environments are not monolithic infrastructures; rather, each is a collection of flexible services and tools that can operate independently from one another, and many separate services are needed to serve the growing community of user groups across different disciplines.

44 When TEXTvre tried to adopt the TextGrid environment, they found that even fundamental tools such the TextGrid XML editor might require reconfiguration, as users prefer to continue to use the environments with which they are familiar. Another major component part of the TEXTvre work was the integration of automated annotation, which is continued today by projects such as Pundit and Korbo,[30] and increasing numbers of tools can be expected in the future. DARIAH has learned from the collaboration experiences of TextGrid and TEXTvre, and to facilitate such communities of tools and services, it is setting up a DARIAH "social marketplace" where they can be sustained, shared, and exchanged.

45 The collaboration between TextGrid and TEXTvre also addressed the creation of a common data infrastructure. Even in this area, where common services might seem easier to achieve than in user-facing tools and services, it was evident that requirements could vary significantly. Moreover, the seemingly straightforward task of integrating a standard digital repository technology into the storage infrastructure of TextGrid has run into difficulties on several levels. Collaboration thus cannot just mean reuse of complex software application but rather needs to function as a genuine exchange of tools, expertise, and experience so that customized solutions that work in different national and institutional environments can be developed and taken forward. Therefore, as far as data infrastructure is concerned, DARIAH has rejected the idea of providing a single data layer, and will instead focus on supporting a range of more diverse services, which are difficult to realize locally as they require a more sustainable environment than project-based funding can provide.

46 TEXTvre has reached the end of its funding, while TextGrid has achieved a third round of funding. During this period, TextGrid is establishing a national working group on the coordination of sustainability and information infrastructure in the humanities (Koordination Geisteswissenschaften Verstetigung/Informationsinfrastruktur [KGW]) with the support of the BMBF. TEXTvre has been taken forward in several new projects, among them the SAWS project and the TTLE service discussed above. While both TextGrid and TEXTvre continue to develop, each has already achieved real (institutional) change in the communities they support.

Bibliography

Blanke, Tobias, Michael Bryant, Mark Hedges, Andreas Aschenbrenner, and Michael Priddy. 2011. "Preparing DARIAH." In *Proceedings: 2011 Seventh IEEE International Conference on eScience*, 158–65. Los Alamitos, CA; Washington, DC: IEEE. doi:10.1109/eScience.2011.30.

Constantopoulos, Panos, Costis Dallas, Petern Doorn, Dimitris Gavrilis, A. Groß, and Georgios Stylianou. 2008. "Preparing DARIAH." In *Digital Heritage: Proceedings of the 14th International Conference on Virtual Systems and MultiMedia, Limassol, Cyprus: Short Papers*, 164–66. Budapest: Archaeolingua. http://pubman.mpdl.mpg.de/pubman/item/escidoc:66841:4/component/escidoc:66842/Preparing_DARIAH.pdf.

Cunningham, Hamish, Diana Maynard, and Kalina Bontcheva. 2011. *Text Processing with GATE*. Sheffield, UK: University of Sheffield Department of Computer Science.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. "GATE: An Architecture for Development of Robust HLT Applications." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 168–75. Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1073083.1073112.

Hedges, Mark, Anna Jordanous, Stuart Dunn, Charlotte Roueché, Marc W. Kuster, Thomas Selig, Michael Bittorf, and Waldemar Artes. 2012. "New Models for Collaborative Textual Scholarship." In *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. Piscataway, NJ: IEEE. doi:10.1109/DEST.2012.6227933.

Jordanous, Anna, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman. 2012. "Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web." In *Proceedings of the 2nd*

International Conference on Web Intelligence, Mining and Semantics, article no. 44. New York: ACM. doi:10.1145/2254129.2254184.

Kuster, Marc Wilhelm, Christoph Ludwig, and Andreas Aschenbrenner. 2007. "TextGrid as a Digital Ecosystem." In *2007 Inaugural IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2007)*, 506–11. Piscataway, NJ: IEEE. doi:10.1109/DEST.2007.372029.

Lagoze, Carl, Sandy Payette, Edwin Shin, and Chris Wilper. 2006. "Fedora: An Architecture for Complex Objects and Their Relationships." *International Journal on Digital Libraries* 6(2): 124–38. doi:10.1007/s00799-005-0130-3.

Neuroth, Heike, Felix Lohmeier, and Kathleen Marie Smith. 2011. "TextGrid – Virtual Research Environment for the Humanities." *International Journal of Digital Curation* 6(2): 222–31. doi:10.2218/ijdc.v6i2.198.

Notes

1 http://www.textgrid.de/en.html
2 http://textvre.cerch.kcl.ac.uk/
3 http://www.dariah.eu/
4 http://digilib.berlios.de/
5 http://cordis.europa.eu/esfri/
6 http://www.d-grid.de/index.php?id=1&L=1
7 http://www.goegrid.de/
8 http://www.earlyenglishlaws.ac.uk/
9 http://www.gasconrolls.org/
10 http://irt.kcl.ac.uk/irt2009/
11 http://epidoc.sourceforge.net/
12 http://www.woerterbuchnetz.de
13 http://collatex.sourceforge.net
14 http://www.fh-worms.de/index.php?id=4616&L=1
15 http://music-encoding.org/
16 Note that it is not possible to apply an XSLT transformation to an XML file being edited with the plugin. The TextGrid team investigated some analogous issues and suggested that it might be possible to remedy them by using virtual workspace folders, which were introduced in a more recent release of the Eclipse plugin. However, this was not high priority for TEXTvre since the King's users were concerned only with TEI file creation.
17 http://lucene.apache.org/solr/
18 http://gate.ac.uk/ie/annie.html
19 https://dev2.dariah.eu/wiki/display/TextGrid/TG-crud
20 http://www.cs.vu.nl/ibis/javagat.html
21 http://www.openrdf.org/
22 http://www.w3.org/TeamSubmission/n3/
23 http://jena.apache.org/
24 http://software.ac.uk/
25 http://www.epsrc.ac.uk/
26 http://www.epcc.ed.ac.uk/
27 https://daasi.de/
28 Funded by HERA (Humanities in the European Research Area).
29 http://www.openannotation.org/spec/core/
30 http://dm2e.eu/first-release-of-pundit-and-korbo-ready-for-testing/

Cite this article

Electronic reference

Mark Hedges, Heike Neuroth, Kathleen M. Smith, Tobias Blanke, Laurent Romary, Marc Küster and Malcolm Illingworth, « TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures

for Textual Scholarship », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 25 June 2013, connection on 10 December 2013. URL : http://jtei.revues.org/774 ; DOI : 10.4000/jtei.774

Authors

Mark Hedges
Mark Hedges is director of the Centre for e-Research and senior lecturer in the Department of Digital Humanities at King's College London, teaching on a variety of modules in the MA programme in Digital Asset and Media Management. His original academic background was in mathematics and philosophy, and he gained a PhD in mathematics at University College London before starting a 17-year career in the software industry, working on large-scale development projects for industrial and commercial clients. He began his career at King's as technical director of the Arts and Humanities Data Service.

Heike Neuroth
Heike Neuroth holds a PhD in Geology and since 1997 she has been working at the Göttingen State and University Library (SUB) in Germany, where she heads the Research and Development Department (RDD). She has been involved in building virtual research environments and research infrastructures for many years. She lectures in information science at the Applied University of Cologne as well as in digital humanities at the University of Würzburg. Additionally, she has been an e-humanities consultant at the Max Planck Digital Library (MPDL) since February 2008.

Kathleen M. Smith
Kathleen M. Smith is a research fellow in the Research and Development Department of the Göttingen State and University Library. She holds an MLS and a PhD in Germanic Literatures, and her research focuses on early modern information networks. Kathleen is currently working on the CENDARI project, which is developing a research infrastructure for the fields of medieval and World War I history.

Tobias Blanke
Tobias Blanke is a senior lecturer in the Centre for e-Research, Department of Digital Humanities, at King's College London. He is a director of DARIAH, a European research infrastructure for arts, humanities, and cultural heritage data, and leads the joint research work for EHRI, a pan-European consortium to build a European Holocaust Research Infrastructure. He holds PhDs in philosophy and in computer science.

Laurent Romary
Laurent Romary is directeur de recherche (director of research) at INRIA, Rocquencourt, France, and guest scientist at Humboldt University in Berlin, Germany. He carries out research on the modeling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He is the chairman of ISO committee TC 37/SC 4 on Language Resource Management, and has been active as member (2001–2007), then chair (2008–2011), of the TEI Council. He currently contributes to the establishment and coordination of the European DARIAH infrastructure.

Marc Küster
Marc Küster is professor for web services and XML technologies at Fachhochschule Worms, Germany. He has a background in physics, literary studies, and history, and his current research interests include the application of information technology in the humanities, Web services, and digital ecosystems. He has served as keynote speaker, program committee member, workshop chair, technical program chair, and general co-chair for various international conferences, and is active in standardization in the fields of cultural diversity, eBusiness, and eGovernment. Since 2008, he has been on leave for a project at the Publication Office of the EU (central data and metadata management).

Malcolm Illingworth
Malcolm Illingworth is an applications consultant at EPCC, one of the foremost supercomputer centers in the world, established in 1990 as a focus for the University of Edinburgh's work in high-performance computing. He works on facilitating the effective exploitation of advanced computing methods to solve real-world problems in academia, industry, and commerce.

Copyright

Abstract

A variety of initiatives for developing virtual research environments, research infrastructures, and cyberinfrastructures have been funded in recent years. The systems produced vary considerably, but they all face the issue of sustainability, namely how to ensure the continued existence of a resource once the project that created it has finished. This paper addresses the sustainability issues faced by the TextGrid and TEXTvre virtual research environments for textual scholarship, examining the inter-project collaboration and cross-fertilization that took place, and investigating how the projects benefited from this exchange. It also examines how their sustainability is being facilitated by the more general-purpose DARIAH infrastructure, and conversely how their existing collaboration can serve as a model for future collaborations within the DARIAH community.

Index terms

Keywords : VRE, ESFRI, tools, services, data architecture, sustainability, research infrastructure