Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions

Tanja Wissik
ACDH-OEAW
Vienna, Austria
tanja.wissik@oeaw.ac.at

Matej Ďurčo
ACDH-OEAW
Vienna, Austria
matej.durco@oeaw.ac.at

Abstract

In this paper we will present an institutional research data workflow model covering the whole lifecycle of the data and showcase the implementation of the model in a specific institutional context. We will present a case study from the Austrian Centre for Digital Humanities, a newly founded research institute for digital humanities of the Austrian Academy of Sciences, which also supports researchers in the humanities as service unit. The main challenge addressed is how to harmonize existing processes and systems in order to reach a clear division of roles and achieve a workable, sustainable workflow in dealing with research data.

1 Introduction[1]

Institutions like universities and academies have an increasing obligation to manage and share research data. For the majority of scholars these endeavours, especially in the humanities, are relatively new and not deeply integrated into their existing working practices: for example, only recently have funding bodies started to request a data management plan which follows open access policies for publications and research data as part of a project proposal[2]. Whereas the traditional non-digital research process consisted only of project planning, data acquisition and data analysis and the publication, in e-research, data sharing, data preservation and data reuse are added to the lifecycle (Briney, 2015).

However, recent studies (e.g. Bauer et al., 2015; Akers and Doty, 2013; Corti et al., 2014) found out, that sharing and reuse of research data is not yet always an integral part of good research practice and that researchers are not familiar with data management plans etc.

A survey carried out in Austria in 2015 (3016 questionnaires) showed significant variations in researchers' data management practice and needs: "Access to self-generated research data by third parties is usually allowed to a limited degree by researchers. While slightly more than half of the respondents stated they allowed access only on request, only one in ten provides their research data as open data for the public; the same number of researchers deny access altogether." (Bauer et al., 2015). The study also reported that 49% of the respondents would need help with project-specific research data management, e.g. creation of data management plan. In a survey study at Emory University in the USA, Akers and Doty (2013) found that "most (~82%) faculty researchers are only somewhat or not at all familiar with requirements for data management or data sharing plans" and "arts and humanities researchers are most likely to be completely unfamiliar with these funding agency requirements for data management plans." A study in the UK in 2008 showed a similar picture: "Only 37% of studied researchers shared their data with collaborators in their own circles and only 20% shared more widely outside of their own network." (Corti et al., 2014: 9). Most concerns about sharing data arise from a lack of knowledge on how to make digital research data sharable for the longer term and a lack of

[2] E.g. Austrian Research Fund (FWF), https://www.fwf.ac.at/en/research-funding/open-access-policy/ (accessed 28.12.2015).

suitable infrastructure in their own institutions. However, knowledge about data management, training and working infrastructure and services are part of successful research data management that has to be adopted by individual researchers as well as institutions like universities and academies. The survey in Austria (Bauer et al., 2015: 11) also found out that "[t]he majority of researchers desire technical infrastructure and project-specific support for research data management. In addition, more than one-third show interest in legal advice, a general help desk, as well as training programs."

This analysis is based on already existing lifecycle or workflow models, taking into account the existing working practice and institutional requirements. Therefore research workflows and the related data management processes vary not only from discipline to discipline but also from one institutional context to another. Once a workflow model is in place, it can serve also as a quality assurance mechanism.

In this paper we will present a case study from the Austrian Centre for Digital Humanities, a newly founded research institute for digital humanities of the Austrian Academy of Sciences, which also supports researchers in the humanities as service unit, where an institutional research data workflow model is being implemented based on already existing lifecycle models.

The context-specific challenge for this undertaking was to bring all the stakeholders together in order to create a model which can simultaneously meet the unique needs of the various sub-disciplines, departments and researchers as well as those of the institute as a whole. Another challenge was being general enough to be applicable to different scenarios including national and international contexts. At the international level the institute is heavily involved in the infrastructure consortium CLARIN-ERIC. One can see the necessity of sound digital management practice at this level, most notably in the institute's role as national coordinator and as service provider of the CLARIN B Centre. This involvement implies a domain specific repository providing depositing services for language resources, the CLARIN Centre Vienna[3]. Furthermore, creating a workflow and data management model that can be applied to the wide variety of different types of data and sources in the arts, humanities and social sciences to be dealt with is a major challenge.

2 Research data lifecycle models

The data lifecycle has becoming an ever more important factor in the researcher's scientific work. This is even more the case given the increasing emphasis on data sharing in research. (Corti et al., 2014). "Life cycle models are shaping the way we study digital information processes. These models represent the life course of a larger system, such as the research process, through a series of sequentially related stages or phases in which information is produced or manipulated." (Humphrey, 2006). They "help to define and illustrate these complex processes visually, making it easier to identify the component parts or distinct stages of the research data" (Carlson, 2014) and the responsible persons or entities. There is a wide range of data lifecycle models, each with a different focus or perspective. The research data lifecycle models can be classified according to the form (linear, circular, non-linear or other models) or (Carlson, 2014) according to the context of the model (individual-based, organisation-based and community-based models) as described by Carlson (2014). In this section, we will present and discuss existing data lifecycle models.

2.1 Models classified according to visualisation form

An example of the linear type is the USGS Science Data Lifecycle Model (see Figure 1). This model describes the data lifecycle from the perspective of research projects and the activities that are performed in phases, e.g. planning, collection, processing, analysis, preservation, publication and sharing of the data for others to reuse. In addition to these activities, there are others that must be performed continually across all phases of the lifecycle, such as the documentation of the workflow process, and the provision of metadata, as well as the backup of data in order to prevent the possibility of physical loss (Faundeen et al., 2013).

[3] http://clarin.oeaw.ac.at

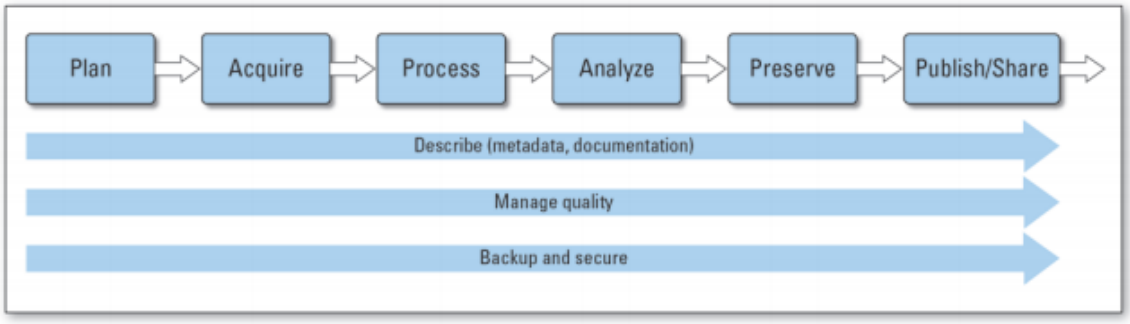

Figure 1: USGS Science Data Lifecycle Model (Faundeen et al., 2013).

There are also circular models which try to reflect the iterative nature of the research process where each step builds on existing material. Circular models seem better suited to describe current research practices increasingly relying on sharing and reuse of data (beyond one researcher or group), an example of which below (Figure 2) shows the e-research and data and information lifecycle (Allan, 2009) with a focus on sharing of data and information.

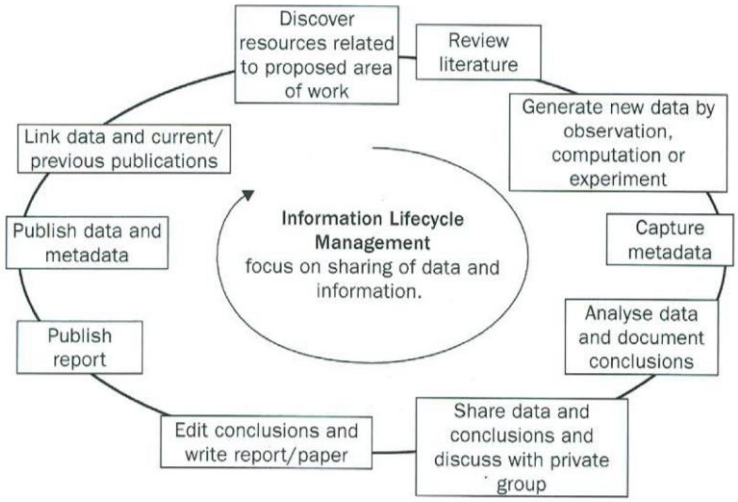

Figure 2: e-research and data and information lifecycle (Allan, 2009).

There are also other types of lifecycle models or workflow models, for example, the non-linear GLBPM model (Barkow et al., 2013) or the OAIS, the Open Archival Information System Reference Model (Lavoie, 2004) which is a concept model for digital repository and archival system. Given that this system does not intend to represent the whole research workflow, it does not fit in the classification above.

2.2 Models classified according to the creator or user of the model

Carlson (2014) describes three different types of life cycle models: individual-based life cycle models, organisation-based models and community-based models. Individual-based models are project-specific (Carlson, 2014) and are often not at an abstract level but contain project related detailed information. Such individual-based models can be helpful in elaborating a data management plan for a specific project. Organisation-based life cycle models "are produced by organi[s]ations offering services or assistance to researchers" (Carlson, 2014). These organisations include universities, libraries, data repositories, publishers etc. An example of an organisation-based model is the University of Oxford Research Data Management Chart (see Figure 3). Compared to individual-based life cycle models, organisation-based life cycle models generalise the different phases of the data lifecycle more since they are not focused on a specific project. From Figure 3, it becomes apparent

that this chart, in contrast to Figure 2, is not organised alongside the research process, but alongside the services that the organisation can offer the researchers in the different stages of the data lifecycle. The University of Oxford has a "Policy on the Management of Research Data and Records". In this policy it is stated, that the university "is responsible for: Providing access to services and facilities for the storage, backup, deposit and retention of research data and records that allow researchers to meet their requirements under this policy and those of the funders of their research; Providing researchers with access to training, support and advice in research data and records management; Providing the necessary resources to those operational units charged with the provision of these services, facilities and training." (University of Oxford, 2014). The model shows that, in compliance with the above mentioned policy, the support offered provides a data management planning checklist as well as services for data backup and data archiving.

Figure 3: University of Oxford Research Data Management Chart (CEOS, 2011).

Models from the third type are called the community-based life cycle models. They have been developed to support the needs of a particular research community and convey recommended best practices in a way that leads to a shared understanding and adoption of these practices in the interested community (Carlson, 2014). An example of a community-based lifecycle model is the DCC Curation Lifecycle Model (Higgins, 2008) (see Figure 4), which describes the different stages of data curation in detail but does not locate the curation process within a research project lifecycle. The model "offers a graphical high-level overview of the lifecycle stages required for successful curation. Generic in nature, the model is indicative rather than exhaustive. When used as an organisational planning tool, it is adaptable to different domains, and extensible to allow curation and preservation activities to be planned at different levels of granularity. It can be used to: define roles and responsibilities; build frameworks of standards and technologies; and ensure that processes and policies are adequately documented. The model identifies: curation actions which are applicable across the whole digital lifecycle; those which need to be undertaken sequentially if curation is to be successful; and those which are undertaken occasionally, as circumstances dictate" (Higgins, 2008). The DDC model (Higgins, 2008) is structured around data (digital objects or databases) and actions. It divides the actions in full lifecycle actions (description and representation information, preservation planning, community watch and participation, curate and preserve), sequential actions (conceptualise, create and receive, appraise and select, ingest, preservation actions, store, access, use and reuse, transform) and occasional actions (dispose, reappraise, migrate).

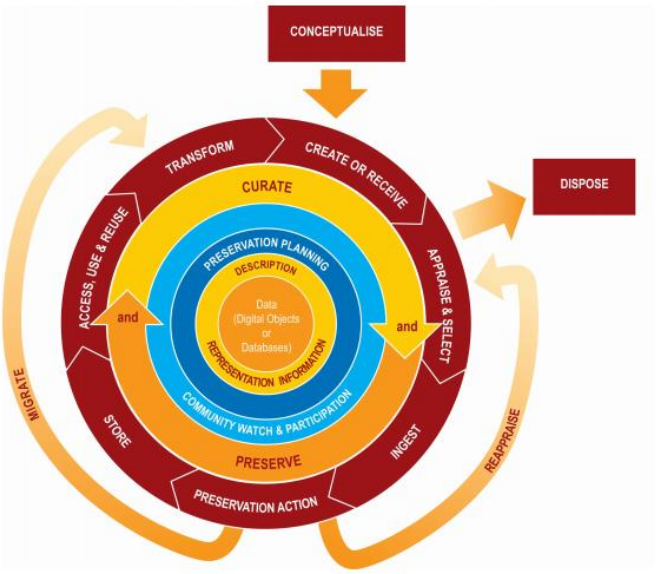

Figure 4: DCC Curation Lifecycle Model (Higgins, 2008).

3 Research Data Management

In this section, we will first define some key terms followed by a description of the institutional case study, the stakeholders and the workflow model. Additionally, we will explain the relation to Clarin and delineate the current status of the implementation.

For this paper, we define research data management as "all data practices, manipulations, enhancements and processes that ensure that research data are of a high quality, are well organized, documented, preserved, sustainable, accessible and reusable" (Corti et al., 2014). Even though the definition of data and research data, especially in the humanities, is subject to intensive discussion (e.g. Sahle, 2015; Kennan and Markauskaite, 2014), it will not be further discussed here. For this paper, we use the definition given by the Consortia Advancing Standards in Research Administration Information (CASRAI)[4].

As mentioned before, data lifecycles are a high level presentation of processes. On the other hand, data management workflows should be specific and detailed enough to serve as blueprint. In order to design the workflow, the stakeholder, the different steps, and their dependences have to be identified for every task/scenario. As Carlson (2014) stated: "Applying life cycle models to support services for managing research data has several benefits". Because "[f]rom its inception to its use and completion, research data will likely undergo multiple transformations in its format, application, use and perhaps even its purpose. Through identifying and naming the transformations that data will undergo as stages in a larger life cycle, organi[s]ations can better target their services […]." (Carlson, 2014).

While the abstract lifecycle models can serve as guidance, they have their limitations. In practice the workflows will usually be more complex with possible variations due to context-specific constraints and because lifecycle models tend to present an idealized version of the processes (Carlson, 2014).

[4] Data: Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records. (http://dictionary.casrai.org/Data) [accessed 30.12.2015]

4 Case study: Institutional Data Management Service at the Austrian Academy of Sciences

There are a lot of publications dealing with institutional case studies, describing the setting up of data management services, as well as the different approaches and challenges encountered (e.g. Choudhury, 2014; Brown and White, 2014; Beitz et al., 2014; Akers et al., 2014; Henry, 2014). When applying data lifecycle models to data management services, different factors have to be taken into account, e.g. who is the target group, what are the best practices and standards in the relevant field or community and what are the real needs of the intended target group (Carlson, 2014). In the following, we will describe the case study of the development and implementation of an institutional data management service at the Austrian Academy of Sciences.

In 2014, the Austrian Academy of Sciences launched the "go!digital"[5] funding initiative supported by the Federal Ministry for Science, Research and Economy (BMWFW) as well as the funding initiative "Digital Humanities: long-term projects on cultural heritage"[6] in order to foster/boost the digital humanities research at the academy and in Austria in general, with special focus on scientific digitisation of cultural heritage as the indispensable base for DH research.

As a natural consequence of this initiative, there has been, as expected, a substantial rise in the amount of new data, leading to a corresponding need to manage the research data and support these projects. In 2015 different institutional stakeholders formed a working group tasked with coordination of the development and implementation of research data management services at the institutional level; these services will be accompanied by technical support, training and workshops on best practice.

4.1 Stakeholders and target group

The following stakeholders are part of the working group *data services*: the Austrian Centre for Digital Humanities (ACDH-OEAW)[7]; the institutional publishing house Academy Press; the institutional computing centre of the Academy (ARZ) and the institutional library (BAS:IS). There are also other stakeholders that are at the moment not part of the working group but nevertheless, play a key role in the data management workflow: third-party service providers for digitisation. The intended target group for the data management services are the researchers[8], both within and outside the academy, especially in the arts, humanities and social sciences. Researchers in life sciences, physics, mathematics etc. use already well established infrastructure for archiving in their relevant fields and are not the main target group.

The ACDH-OEAW runs a domain specific repository for the arts and humanities, with a particular emphasis on language resources, for which we operate the Language Resources Portal which is part of the CLARIN Centre Vienna (*CCV/LRP*)[9]. The ACDH-OEAW also offers a range of applications and services for processing, analysing, visualising and querying different kinds of data.

The Press has been operating the institutional repository of the Academy, *epub.oeaw*[10] that is designated to hold primarily scientific publications, but increasingly also research data. The repository serves a double role: publication and archiving, data in the repository being replicated to the Austrian National Library (Stöger et al., 2012). So, while there is some overlap in the task description of *epub.oeaw* and *CCV/LRP*, there are distinct features, that justify the co-existence of the two repositories.

Currently, the stakeholders are elaborating a common strategy to act as a coordinated network of providers for data-related services, with clear division of roles. In this plan, ACDH-OEAW will concentrate more on the interaction with the researchers (consulting, data modelling), development

[5] http://www.oeaw.ac.at/en/fellowship-funding/promotional-programmes/godigital/

[6] http://www.oeaw.ac.at/en/fellowship-funding/promotional-programmes/digital-humanities-long-term-projects-on-cultural-heritage/

[7] http://www.oeaw.ac.at/acdh

[8] In this paper, as researchers we mean research staff of the Austrian Academy of Sciences as well as non-members of the Austrian Academy of Sciences who are conducting research in collaboration with the Academy or are making use of the offered services and are willing to deposit data in one of the described repositories.

[9] https://clarin.oeaw.ac.at/

[10] http://epub.oeaw.ac.at/

and provision of tools for processing, analysing, visualising the data. The Press will keep running the repository for archiving and publishing of publications and certain types of research data. However, not all kinds of resources are equally well suited for the digital asset management system underlying *epub.oeaw*, particular examples of which are: relational databases, corpora and graph-based data. Thus, the working group still needs to work a strategy for archiving for this kind of data. Furthermore, there are plans to establish in-house capacities for digitisation at the institutional library that also serves as an important content provider.
One of the challenges was to bring all the stakeholders together and to develop a common strategy how to deliver a data management service together, since these stakeholders haven't worked together until recently. One of the peculiarities of the present case study is that in contrast to the usual setup, where the institutional libraries are the driving forces in the process and deliver most of the services related to data management (Choudhury, 2014; Brown and White, 2014; Beitz et al., 2014; Akers et al., 2014; Henry, 2014), in our case the coordinating unit, the Austrian Centre for Digital Humanities, is a research institute that also functions as service unit, and therefore, the institute is involved also as research partner.
In the following section we will explain the workflow model.

4.2 Workflow Model

In Figure 5 the proposed research data management workflow is illustrated from the perspective of the institute, the ACDH-OEAW with a focus on projects from the arts, humanities and social sciences. The key roles in this model are taken by the researcher, the institute, the publishing house, the library and third party service provider. The institutional computing centre of the Academy (ARZ) is not present in the model, however it is still an indispensable partner as it runs the basic technical infrastructure (servers, storage, networks, etc.). If in Figure 5 for one task only one form is visible, then only one stakeholder is responsible for this task, if there are more overlapping forms in different grey tonalities, black or white then the responsibilities are shared.

In this model below, six different phases are shown which are as follows: the pre-processing (divided into proposal stage and granted stage), the processing, the storage, the publishing and the reuse phase as well as quality assurance. As shown in the model (Figure 5), not all the phases are clear-cut and they can overlap. The quality assurance process is special, as it accompanies and underlies the whole workflow. There are mainly two different scenarios to which the institutional research data management model has to be applied. The first scenario is when a new project proposal is written, here we call this scenario *new project* (Figure 5) the second is when the project is already over, here (Figure 5) we call this scenario *legacy data*. In the following we will describe the two scenarios in detail.

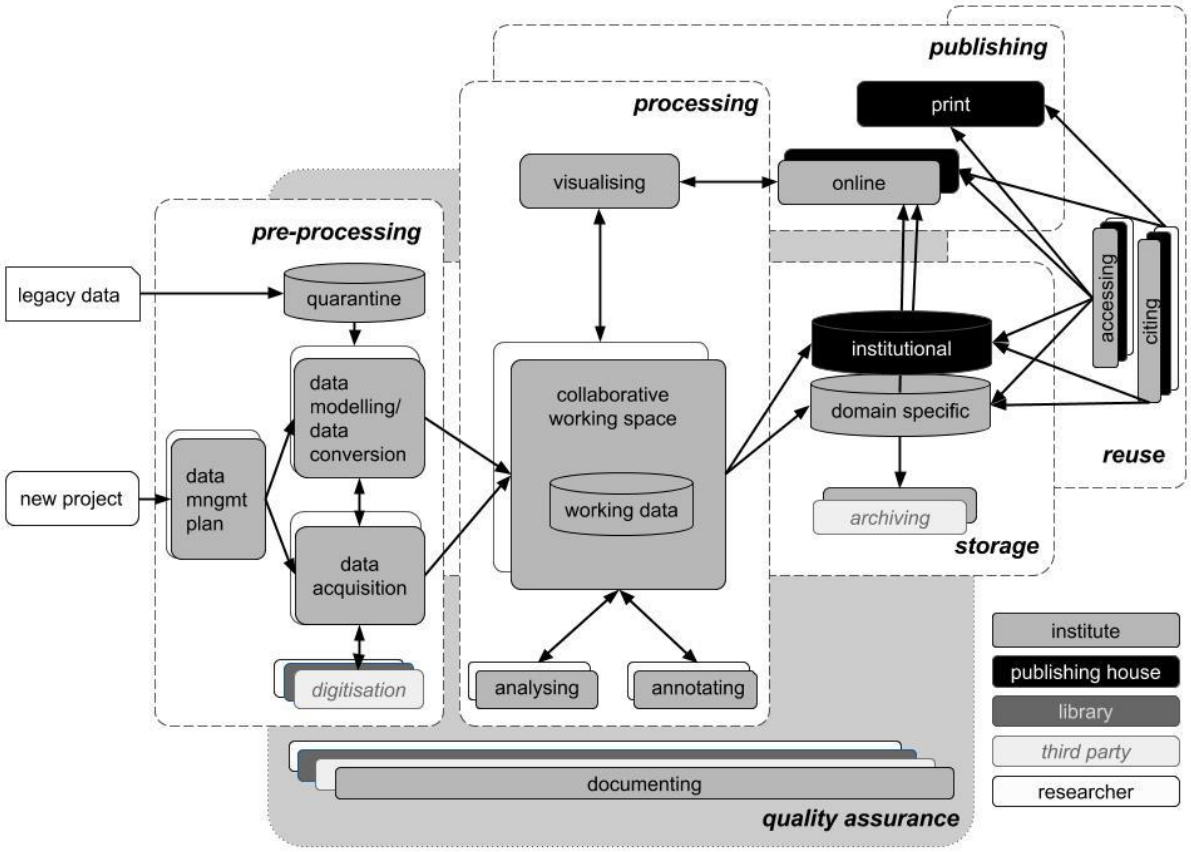

Figure 5: Proposed institutional research data management workflow.

4.3 First scenario: new project

In this first scenario, the researcher approaches the institute (as part of the *data services* team) for advice with a project idea in the proposal phase. There, the new project enters the pre-processing phase which itself has two different stages, the *proposal* and *granted* stage. The first step in the proposal stage is the elaboration of the data management plan that most of the funding agencies nowadays require for a grant application. The institute advises the researcher on data management issues, especially on the resources (people, equipment, infrastructure and tools) that have to be taken into account in the project budget. In the ideal case, the institute and the *data services* group is included into the project proposal. If the project gets funded, the project enters the *granted* stage at which time, the data collection starts. If the new project involves digitisation, this is also part of the pre-processing phase and is either done by the researchers themselves or by a third party service provider.[11] In parallel to collecting or acquiring the data, the institute elaborates the data model together with the researcher. As data model we understand "[a] model that specifies the structure or schema of a dataset. The model provides a documented description of the data and thus is an instance of metadata"[12] as defined by the Data Foundation and Terminology Working group of the Research Data Alliance (RDA). Based on the data model and the requirements of the project, formats for data and metadata are discussed and chosen in accordance with best practices and standards in order to avoid data loss and conversion problems in the future. If we compare our model with the previous discussed lifecycle models, the pre-processing phase in our model corresponds to the activities *plan* and *acquire* in the USGS Science Data Lifecycle Model (Figure 1) or to *generate new data* and *acquire metadata* in the model of Allen (2009) in Figure 2. In the model in Figure 3 it would

[11] Figure 5 illustrates that the Academy library also offers digitisation services. This service is not yet in place but it is expected to be enacted sometime this year.

[12] RDA Term Tool, entry "data model" available at http://smw-rda.esc.rzg.mpg.de/index.php/Data_Model ([accessed: 30.12.2015]

correspond to *data management planning* and in the DDC Model (Figure 4) it would correspond to *conceptualise* and *create.*

After the pre-processing phase the data enters the processing phase during which all research activities related to previously acquired data take place. In referring to *processing*, we specifically mean: "performing a series of actions in something (an input) in order to achieve a particular result (output)."[13] Some of the actions are mentioned in the model (analysing, annotating, visualising), but they are not exhaustive. If we take the NeDiMAH Methods Ontology as a reference point, annotating would be a subtype of analysing, but we decided to depict them at the same level, given the importance of the annotation step in the research process. Ideally, the researchers work in an integrated collaborative working space, where they get offered a series of tools for annotating, analysing, visualizing etc., run as a service by the data services working group. Data visualisation is helpful in detecting patterns and performing analysis, and therefore it is used in the collaborative working space during the processing phase and it is used in the publication phase for online publication of the data. In the model the visualising activity is in the overlapping of the processing and the publishing phase in order to reflect these two purposes. Currently the above mentioned portfolio of tools is being built up combining existing open source applications as well as specific solutions to a task. Thanks to the strong international involvement of ACDH-OEAW, the tool development is deeply embedded in the activities of the research infrastructures CLARIN & DARIAH as well as RI projects, most prominently the new H2020 project PARTHENOS[14]. The processing phase corresponds to the activities *process* and *analyse* in the USGS Science Data Lifecycle Model. The collaborative working space reflects the activities *analyse data and document conclusions* and *share data and conclusions and discuss with private group* in the data lifecycle by Allan (2009). In both lifecycle models, publishing activities are foreseen as well as in our proposed workflow.

An important activity, especially in relation to future reuse (Corti et al., 2014) of data, is documenting. Documenting is understood as "providing information regarding each and every step of the activities that took place in a project, in order to describe how everything was done and enable someone that was not initially involved to understand."[15] Data documentation includes information on data creation, content, structure, coding, anonymization etc. There are two types of documentation, the high level description, also known as study-level documentation and the data level documentation (Ibid). If we have a closer look at the model, the documenting can be found as part of the quality assurance, that runs alongside all the processes Already in the data acquisition and digitisation, documenting plays an important role in order to achieve reusable data at the end of the workflow.

It is important to underline that all the phases as well as the whole workflow cannot be seen as a simple step or linear sequence of steps, but rather a complex, non-linear, iterative process, both within one project as well as beyond the project boundaries

In the storage phase, underlying the whole workflow, the data and metadata are stored and archived. We need to distinguish different kinds of storage. In the pre-processing phase during the data collection, large amounts of data is produced that is the starting point/serves as base for the whole further process and needs to be secured and made accessible within the workspace. In the processing phase, a lot of additional data is produced, oftentimes of transitional nature. We call this "working data". Stable data – raw captured data as well as secondary data / enrichments contributed in the processing phase – aimed at long-term availability and/or publication is moved to the institutional or domain specific repository, which in the long run represents the main source for the datasets. Before the data will be ingested in one of the repositories, licence issues have to be discussed and agreements have to be signed. At the archiving stage, it is necessary to ensure long-term availability of the data even beyond a disaster scenario e.g. main repository is damaged through fire or similar. This involves geographically distributed replication/mirroring of the data to reliable providers of storage services, like scientific data centres. The data from the repository *epub.oeaw* is already being replicated to the Austrian National Library. Additionally, we build up alliances with national providers as well as

[13] Definition taken from the NeDiMAH Methods Ontology (NeMO) available at http://nemo.dcu.gr/index.php?p=hom [accessed: 30.12.2015].

[14] http://www.parthenos-project.eu/

[15] Definition taken from the NeDiMAH Methods Ontology (NeMO) available at http://nemo.dcu.gr/index.php?p=hom [accessed: 30.12.2015].

international players mainly in the context of the EUDAT initiative. Archiving and preservation activities are also mentioned in the USGS Model, in the Oxford Research Data Management Chart and in the DCC Model.
The publishing phase refers primarily to presentation, online and/or print, of the results of the project but also – in line with the open access policy and subject to copyright and ethical restriction – the provision of the underlying research data. Enabling discoverability and citability of the research data is a precondition for effective reuse. The institute and publishing house are providing infrastructure and user interfaces for researchers to search for data and publications and to access them e.g. via the interface of *epub.oeaw* (Stöger et al., 2012). Next to direct access to the data, it is crucial to ensure wide-spread dissemination of the data, again ensured by the combined competencies of Press, library and ACDH-OEAW. While Press ensures indexing of the resources by services like Google Scholar and OpenAIRE, ACDH-OEAW pushes into the more domain-specific channels in the context of CLARIN and DARIAH. One important issue in the reuse phase is proper citation. Proper citation of publications, in the humanities especially of print publications, is an integral part of good research practices. But not all the researchers in the humanities are yet familiar with citations of primary or secondary data sources or data sets or the citation of digital editions. One increasingly popular possibility to help researchers is to integrate citation recommendation within the online presentation of the resources[16]. For data sets the attribution of a unique persistent identifier is essential. While there are several standard persistent identifier (PID) systems (see Corti et al., 2014; Briney, 2015) so far the most relevant to the Academy are Digital Object Identifiers (DOI).The institutional repository *epub.oeaw* is assigning DOIs to each uploaded research result (Stöger et al., 2012). In LRP every resource is assigned a handle-based[17] PIDs in accordance to CLARIN requirements. However, it is essential to use the persistent identifier in the citation, because it helps tracking data citations (Briney, 2015) and use recommended formats of data citations, e.g. Starr and Gastl (2011) resembling traditional print publication citations.

4.4 Second scenario: legacy data

The second scenario, covered by the workflow, is the so called *legacy data* scenario. As legacy data we understand data that fall into the category of dark data or at-risk data. More often, we deal with at-risk data, that is data that are at risk of being lost due to the fact that the project is already over, and the stored data is not well or not at all documented (including missing metadata or the data has been detached from supporting data or metadata) and therefore not useable or reusable or it is stored on a medium that is obsolete or at risk of deterioration.[18]
When confronted with legacy data, in a first step, all the relevant data is stored, as shown in Figure 5, in a kind of "quarantine" repository to be further processed. Then the data and the data model/structure are examined, especially with respect to the suitability of the format, existence of metadata and documentation and internal structure of the data. Based on the analysis, it is decided if the data has to be converted and the data model needs to be adapted, transformed together with the estimation of the required resources of such transformation. Then the data is stored (see storage phase above) in the repositories and archived without going through the processing phase. Usually, there are only limited resources to deal with legacy data, the primary goal is to ensure a reliable deposition of the data and the accessibility for other researchers. Thus as long as no new user/project interested in this data arises, no interaction with the data is expected in the working space, nor is an online publication.

[16] E.g. in the ABaC:us – Austrian Baroque Corpus digital edition a citation suggestion is generated with each query: Abraham â Sancta Clara: Todten-Capelle. Würzburg, 1710. (Digitale Ausgabe) Vorrede [S. 14]. In: ABaC:us – Austrian Baroque Corpus. Hrsg. von Claudia Resch und Ulrike Czeitschner. <https://acdh.oeaw.ac.at/abacus/get/abacus.3_48> abgerufen am 3. 1. 2016
[17] http://www.handle.net/
[18] Modified definition taken from CASRAI Dictionary: legacy data available at http://dictionary.casrai.org/Legacy_data [accessed 07.03.2015]; dark data available at http://dictionary.casrai.org/Dark_data [accessed 07.03.2015]; at-risk data available at http://dictionary.casrai.org/At-risk_data [accessed 07.03.2015]

4.5 Relation to CLARIN

As mentioned before, the development or adaptation of an institutional-based model should take into account the relevant best practices and standards in the community of the intended target group. Given that ACDH-OEAW runs a CLARIN Centre[19] and is a national coordinator of CLARIN activities, many aspects of the workflow are strongly guided by the requirements expected by CLARIN-ERIC[20] – assignment of persistent identifiers, metadata in CMDI (Component Metadata Infrastructure) format (Broeder et al. 2010), OAI-PMH[21] (Open Archives Initiative Protocol for Metadata Harvesting) endpoint as a dissemination channel for the metadata harvested by the CLARIN harvester. One of the aims of the presented strategy is to make new resources automatically available via the CLARIN infrastructure.

Currently, for the resources we use 4 different CMDI profiles, and make the resources available in different forms, partly as raw data, partly within complex web applications that allow search and browsing through the data via different dimensions (linguistic, semantic). These steps are related to the reuse phase in the research data management model in Figure 5. The access to resources and services is managed through Federated Identity.

In 2016, CCV/LRP is scheduled for reassessment. In preparation for this, the repository solution will be overhauled, taking into account lessons learned in the last two years, aiming for tighter integration with the institutional repository *epub.oeaw* (eliminating redundancies). As part of this process, language resources already existing in the *epub.oeaw* repository shall be made accessible within CLARIN (primarily by providing appropriate CMDI records). One central challenge in this task will be to reflect the broader role that the ACDH-OEAW has lately assumed covering not just language resources but expanding to a broad spectrum of disciplines in the context of digital humanities (archaeology, history, art history, etc.). Here we aim – in accordance with the principles of the research infrastructures – for a setup with common/harmonized technical infrastructure in combination with domain- or project-specific solutions/views building on top of it.

With respect to the tools offered for use, there is a reciprocal relation to CLARIN, where tools from the CLARIN community are part of the portfolio, like *WebLicht* (Hinrichs et al. 2010) as well the solutions developed at the institute are made available to the whole CLARIN community, like the SMC Browser (Durco, 2013), Vienna Lexicographic Editor (Budin et al., 2013), or the corpus shell[22] framework

With respect to long-term archiving we plan to take advantage of the relation of CLARIN-ERIC to the EUDAT initiative.

4.6 Current status

Currently, the model is being implemented. Many parts/components of the model are already available (like the repositories, individual processing and visualisation tools, the publishing workflow), the main task in 2016 will be to provide the glue between these components by establishing the procedures inside the *data services* working group and make the services accessible/usable by the target audience – the researchers of the academy and of the broader Austrian DH community.

The usefulness and appropriateness is currently being tested on a number of research projects, especially from the calls *go!digital* and *Digital Humanities: long-term projects on cultural heritage*, all of which started last year. A few examples of types of projects and data we are dealing with include APIS project[23], which aims to enrich and convert a large biographical lexicon with the help of NLP tools into richly structured Linked Open Data. These data will then be made available for exploration through appropriate interactive visualisation means; similarly in *exploreAT!*[24] huge amounts of heterogeneous data gathered over more than a hundred years and available in different digitisation stages and formats will be harmonized (adhering to the LOD paradigm) and made available online,

[19] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-105
[20] hdl:1839/00-DOCS.CLARIN.EU-77
[21] http://www.openarchives.org/pmh/
[22] https://clarin.oeaw.ac.at/corpus_shell
[23] Austrian Prosopographic Information System - http://www.oeaw.ac.at/acdh/apis
[24] http://www.oeaw.ac.at/acdh/exploreat

both as raw structured data and as rich explorative applications; in DEFC [25] a database of archaeological sites and finds is being developed;[26] aims at full linguistic and semantic enrichment of the historic texts of Baedeker using TEI/XML as the native format. These are just four out of a number of projects to sketch the variety of data and requirements the *data services* team is confronted with. From historic manuscripts to archaeological sites and finds - with each project we learn something new and update our data management workflow and with each new project we learn more about how best to manage, store and present online data from the humanities, arts and social sciences. Alongside the testing of the model, we also give training workshops in order to raise awareness and understanding and to improve research data management skills.

5 Conclusion and outlook

In this paper we presented an institutional workflow model for research data as it is currently being implemented at the Austrian Academy of Sciences, coordinated by the ACDH-OEAW, a newly founded research institute of the Academy that acts also as a service unit for researchers in the art and humanities in the institutional and national context. Starting from abstract (research) data lifecycle models, we discussed the stakeholders and scenarios for the specific institutional settings and elaborated a workflow model that caters to the specific situation of the Academy.
Just like Higgins (2008) stated that the DCC Model "is not definitive and will undoubtedly evolve", also the ACDH-OEAW model will evolve. Even once a service is fully functional, the evolution of data-dependent research practices and the changing research technologies have to be monitored in order to adapt the service to changing demands.
The paper shows that the elaboration of an institutional research data workflow model is important since there is no "one-size-fits-all-solution", e.g. Higgins (2008) mentioned "domain-specific variations" of the DCC model, but high level data lifecycle models are a good basis to start with and to adapt to the specific institutional context. The elaboration or adaptation of already existing models depends on different aspects like target group, relevant best practices and standards and real world needs of the intended target group. Once the workflow model is implemented, it can not only be used as quality assurance measure but it can also guide the researchers in the project planning phase, when and whom to approach for advice, assistance and support.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions on this paper.

References

[Allan, 2009] Robert Allan. 2009. Virtual Research Environments. From portals to science gateways. Chandos Publishing, Oxford, UK.

[Akers et al. 2014] Katherine G. Akers, F.C. Sferdean, Natsuko H. Nicholls, Jennifer A. Green. 2014. Building Support for Research Data Management: Biographies of Eight Research Universities. *International Journal of Digital Curation*. 9(2):171-191.

[Akers and Doty, 2013] Katherine G. Akers and Jennifer Doty. 2013. Disciplinary Differences in Faculty Research Data Management Practices and Perspectives. *The International Journal of Digital Curation*, 8(2):5-26.

[Barkow et al., 2013] Ingo Barkow, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, Wolfgang Zenk-Möltgen. 2013. *Generic longitudinal business process model. DDI Working Paper Series – Longitudinal Best Practice, No. 5.* DOI: http://dx.doi.org/10.3886/DDILongitudinal2-01

[Beitz et al., 2014] Antony Beitz, David Groenewegen, Cathrine Harboe-Ree, Wilna Macmillan and Sam Searle. 2014. Monash University, a strategic approach. In: Graham Pryor, Sarah Johnes and Angus Whyte. 2014.

[25] Digitizing Early Farming Cultures - http://www.oeaw.ac.at/acdh/defc
[26] https://acdh.oeaw.ac.at/acdh/traveldigital

Delivering Research Data Management Services. Fundamentals of good practices. Facet Publishing, London, UK. 163-189.

[Briney, 2015] Kristin Briney. 2015. Data Management for Researchers. Organize, maintain and share your data for research success. Pelagic Publishing, Exeter, UK.

[Bauer et al., 2015] Bruno Bauer, Andreas Ferus, Juan Gorraiz, Veronika Gründhammer, Christian Gumpenberger, Nikolaus Maly, Johannes Michael Mühlegger, José Luis Preza;, Barbara Sánchez Solís, Nora Schmidt andChristian Steineder (2015): Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung. Report 2015. Version 1.2. DOI: 10.5281/zenodo.32043. Online available at http://phaidra.univie.ac.at/o:407513

[Broeder et al. 2010] Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Philip Withers, Peter Wittenburgand Claus Zinn.2010. A data category registry-and component-based metadata framework. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*. European Language Resources Association (ELRA). 43-47.

[Brown and White, 2014] Mark L. Brown and Wendy White. 2014. University of Southampton – a partnership approach to research data management. In: Graham Pryor, Sarah Johnes and Angus Whyte. 2014. Delivering Research Data Management Services. Fundamentals of good practices. Facet Publishing, London, UK. 135-161.

[Budin et al, 2013] Gerhard Budin, Karlheinz Moerth and Matej Durco. 2013. European Lexicography Infrastructure Components. In Iztok Kosem, JelenaKallas, Polona Gantar, Simon Krek, Margit Langemets and Maria Tuulik (eds.), Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013 (pp. 76–92). Tallin, Estonia: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. http://eki.ee/elex2013/conf-proceedings/

[Carlson, 2014] Jake Carlson. 2014. The Use of Life Cycle Models in Developing and Supporting Data Services. In: Joyce M Ray (ed). 2014. *Research Data Management. Practical Strategies for Information Professionals*. Purdue University Press, West Lafayette, IN. 63-86.

[CEOS, 2011] CEOS. 2011. *CEOS Working Group on Information Systems and Services. Data Life Cycle Models and Concepts. CEOS Version 1.0*. Available at http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS_DSIG-Data-Lifecycle-Models-and-Concepts-v8_Sep2011.docx

[Choudhury, 2014] G. Sayeed Choudhury. 2014. John Hopkins University Data Management Services. In: Graham Pryor, Sarah Jones and Angus Whyte. 2014. *Delivering Research Data Management Services. Fundamentals of good practices.* Facet Publishing, London, UK: 115-133.

[Corti et al., 2014] Louise Corti, Veerle Van den Eynden, Libby Bishop and Mattew Woolard. 2014. *Managing and sharing research data. A guide to good practice*. SAGE, London, UK.

[Durco, 2013] Matej Durco. 2013. *SMC4LRT - Semantic Mapping Component for Language Resources and Technology*. Technical University, Vienna, Austria. http://permalink.obvsg.at/AC11178534

[Faundeen et al., 2013] John L. Faundeen, Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, Cassandra C. Ladino, Steven Tessler, and Lisa S. Zolly 2013. *The United States Geological Survey Science Data Lifecycle Model. U.S. Geological Survey Open-File Report 2013–1265*, 4 p, http://dx.doi.org/10.3133/ofr20131265.

[Henry, 2014] Geneva Henry. 2014. Data Curation for the Humanities. Perspectives From Rice University. In: Ray, Joyce M. (ed). 2014. *Research Data Management. Practical Strategies for Information Professionals*. Purdue University Press, West Lafayette, IN. 347-374.

[Higgins, 2008] Sarah, Higgins. 2008. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation.* 3(1):134-140.

[Hinrichs et al., 2010] Marie Hinrichs, Thomas Zastrow and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. Paper presented at LREC 2010, Valetta, MT.

[Humphrey, 2006] Charles Humphrey.2006. E-science and the life cycle of research. Available at http://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/bakgrunn/life-cycle.pdf

[Kennan and Markauskaite, 2015] Kennan, Mary Anne and Lina Markauskaite. 2015. Research Data Management Practices: A Snapshot in Time. *International Journal of Data Curation*. 10(2):69-95.

[Lavoie, 2004] Brian F. Lavoie. 2004. *The Open Archival Information System Reference Model: Introductory Guide*. OCLC Online Computer Library Center.

[ICPSR, 2012] Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI. Available at http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

[Sahle, 2015] Patrick Sahle. 2015. Forschungsdaten in den Geisteswissenschaften. *SAGW Bulletin*, 2015(4)4(2015):43-45.

[Starr and Gastl, 2011] Starr, Joan and Angela Gastl. 2011. A Metadata Scheme for DataCite. *D-Lib Magazin*, 17(1/2). doi:10.1045/january2011-starr

[Stöger et al., 2012] Herwig Stöger, Vittorio Muth - Georg Lasinger. 2012. *epub.oeaw Benutzerhandbuch. Das Publikationsportal der Österreichischen Akademie der Wissenschaften*. Österreichische Akademie der Wissenschaften. Vienna, AT. Available at http://epub.oeaw.ac.at/dokumentation14/0000_Epub.UserGuide_1.4_printable.pdf

[University of Oxford, 2014] University of Oxford. 2014. *Policy on the Management of Research Data and Records*. Oxford, UK. Available at http://researchdata.ox.ac.uk/files/2014/01/Policy_on_the_Management_of_Research_Data_and_Records.pdf