IJDC | *General Article*

Curating Humanities Research Data: Managing Workflows for Adjusting a Repository Framework

Hagen Peukert
University of Hamburg

Abstract

Handling heterogeneous data, subject to minimal costs, can be perceived as a classic management problem. The approach at hand applies established managerial theorizing to the field of data curation. It is argued, however, that data curation cannot merely be treated as a standard case of applying management theory in a traditional sense. Rather, the practice of curating humanities research data, the specifications and adjustments of the model suggested here reveal an intertwined process, in which knowledge of both strategic management and solid information technology have to be considered. Thus, suggestions on the strategic positioning of research data, which can be used as an analytical tool to understand the proposed workflow mechanisms, and the definition of workflow modules, which can be flexibly used in designing new standard workflows to configure research data repositories, are put forward.

Received 20 October 2016 ~ *Revision received* 11th January 2017 ~ *Accepted* 11th January 2017

Correspondence should be addressed to Hagen Peukert, Monetastraße 4, 20146, Hamburg, Germany. Email: hagen.peukert@uni-hamburg.de

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v12i2.571
DOI: 10.2218/ijdc.v12i2.571

Introduction

Albeit the common name ‘humanities’ suggests otherwise, the humanities reveal probably the most diverse mix of data in the entire academic faculty. Research data collected in the humanities reveal a tremendous degree of heterogeneity ranging from mere texts in written, spoken, transcribed, or otherwise enriched forms by glosses or handwritten markings, to formal and informal proofs, test series, musical scores, archaeological 3D-models, and complex, multi-layered audio-visual annotated corpus collections.

Due the managerial levels of the university administration seeing the advantages of sustainable data, extensible applications and data/application reuse, data centers have been established that are specifically designed to curate research data. However, these are equipped with strict budget constraints. Indeed, these data centers are deployed on grounds of organizational structure and not, as one would expect, on grounds of the needs, size, and structure of the data to be curated. This bears important consequences to the curation of research data in the humanistic disciplines.

The theoretic underpinnings of the model to be introduced here originate in research and case studies in Process Management and Strategic Management. Of course, the specific arrangement to define the very particular nature of an otherwise too general model cannot be found anywhere near Management Studies, but needs the knowledge of experienced data scientists. In analogy to successful marketers faced with heterogeneous products and trying to tidy up a new as well as confusing market situation with the aim to know how to approach customers and serve their specific needs, data curators in the humanities need to identify the different requirements of the users of data, how best to serve them and align these to the characteristics of the data and models. Moreover, data workflow management is supposed to provide instruments that help to understand the processes and analyze curation tasks to find an optimal workflow at given cost targets.

This paper gives a detailed account of a strategic approach to data curation workflow management. In the first section the theoretical background, i.e. the derivation of the model and its adjustment to research data is described. In the two sections to follow, the results in the form of the strategic positioning as a portfolio technique and the definition of workflow modules are presented. The fourth section clears advantages and disadvantages and presents possible extensions of the suggested model.

Theoretical Background

Using Strategic Management Methods in Data Curation Management

Research in Strategic Management, and more particularly in Strategic Marketing, has been carried out systematically over the past 60-odd years. As the etymology of the term strategy suggests, Management Studies have borrowed heavily from Military Science (von Clausewitz, 2012, 2011; Sun, 2007). Modern market entry strategies still bear names like frontal attack, flank attack, or guerrilla. Yet, the underlining meaning of strategy can still be expressed in more general terms as a long-term behavioral plan to achieve main goals in due consideration of the interaction with the environment (Thompson and Strickland, 1992; Jauch and Glueck, 1988; Hünerberg, 1994). Thus, it is adaptable to a broad range of organizational problems of planning and optimizing.

It is a striking fact that, by and large, the early theoretical constructs in Strategic Management have withstood the ravages of time. They have been refined, empirically reproved, adjusted, and extended, but the core idea has proven to be robust over the decades (Ghemawat, 2000; Nag, Hambrick and Chen, 2007). Findings of classic papers presenting generic strategies (Ansoff, 1965), competitive strategies (Porter, 1980; Kotler and Armstrong, 2016), or portfolio-analyses (Boston Consulting Group, 1973) are still used as strategic tools in practice in global enterprises as well as successful small businesses. Concepts such as lifecycle analyses and experience curve analysis (effects of scale, scope, savings, and learning) have already made its way to the management inventory of data scientists. So it seems justified to apply another part of managerial theory to the fairly recent field of data management and profit from the about 50 years of experience in Strategic Management when aligning its concepts to a very peculiar product, i.e. humanities research data.

The above argument does not imply that the literature is straightforward and well arranged and it also does not suggest blindly applying the entire body of strategic business instruments and methods to data curation management. It needs profound knowledge of the product, data, and the requirements of its customers, users, to make the right decisions about what to select. Yet, the more abstract procedure of how to go about developing an appropriate strategy is recognized as a general truth here.

Standardization, Positioning, and Timing as Strategies in Data Curation

To shed some light on the diverse marketing strategies Hünerberg (1994) following Dahringer and Mühlbacher (1991), Keegan (1989) and Segler (1986) proposes a simple classification of two basic strains of strategic decisions: fundamental and instrumental. Fundamental strategies comprise strategies of market selection, market behavior, and market implantation – all of which more or less depend on the competitive situation. Instrumental strategies aim at solutions to questions of standardization, positioning, and timing. While the entire set of fundamental strategies is simply not applicable to data management for missing a competitive component in the current situation, the instrumental strategies show obvious parallels to what has to be managed in data curation workflows and thus they are worth further exploration. Although the theory purports that the choice of instrumental strategies is directly derived from its fundamental strategies, it is the advantage in data management not to care about competitive analyses as presumed in fundamental strategies and take it as a given. Consequently, a closer look will be taken at standardization, positioning, and timing only.

Timing in Strategic Management means the specification at which point in time, how frequent and at which duration the marketing instruments have to be used (Dahringer and Mühlbacher, 1991; Kreutzer, 1989). Applied to data management and data curation, it would mean at which point in time the curation process should be started, how long it should take and in which intervals data curation recurs. On a more abstract level, the data curator has the choice between two strategies, waterfall and sprinkler, and their combination. Waterfall strategies suggest a succession in the workflow of each single project so that each data project is curated from the beginning to the end before starting a new data curation project. A sprinkler strategy would tackle all available projects at a time. If the duration for finalizing each project is to be kept constant, it is clear that more human resources are required for a sprinkler strategy, since the latter is faster by a factor equal to the number of projects and there are probably no curation workflows whose economies of scale and scope could compensate such factor even approximately. Nevertheless, there are saving effects in the duration of curating data if several projects can be classified on grounds of similar features. Similar projects could then be curated in specially adjusted workflows at the same time. This relates to a combination of waterfall and sprinkler strategies.

The ability to classify data projects is most relevant for the other two strategic decisions that data curators should be concerned about: standardization and positioning. In fact, an appropriate classification is the decisive criteria on the question of to what degree a data curation workflow can be standardized. As a general rule, the more standardization is possible the better and the lower the curation costs. Regarding the question of standardization, there is little deviation in analyzing products in consumer markets and data projects. In contrast, the concept of strategic positioning needs substantial adjustment since it is not the behavior of the market that defines the typological space, in which the data projects are to be evaluated. Strategic positioning in Research Data Management could be placed within the dimensions of data sustainability, cost of data curation, data accessibility, or data usability. Which dimensions are most appropriate can be studied with a suitable procedure that is provided by strategic management expertise.

Portfolio Analysis as an Apt Method of Strategy Development in Data Curation

The literature in Strategic Management enlists about a dozen procedures that support arriving at an effective strategy. Prominent examples are gap analyses, experience curves, product lifecycle analysis, portfolio techniques, situation analysis such as SWOT (strength, weaknesses, opportunities, threads), prognostic methods such scenario analysis, PIMS (Profit Impact of Market Strategies), or balanced scorecards. From these, portfolio analysis is known to be easiest to apply as an analytic tool to various other contexts outside the realm of business studies. Furthermore, together with lifecycle analysis, experience curves and scenario analysis, portfolio techniques do not necessarily absorb competitive advantages, which do clearly not correspond to the situation in data curation management nowadays, and therefore portfolios can be easily detached from incorporating information on markets and competitors. Last, lifecycle analysis and experience curves, both of which are already known to data scientists, can be combined with portfolio analysis to arrive at a more complete understanding of the overall data curation process design. In brief, portfolio techniques seem to be a good candidate to help in planning and organizing workflows in data curation.

A portfolio strategy is a method of analyzing objects of interest in terms of their contribution to strategic goals (Bartol and Martin, 1998). To understand what portfolio analysis does, it is probably best to look at the method itself. The original method of portfolio analysis prescribes five steps.

1. Defining strategic business units (SBU)
2. Consolidating determinants to two dimensions (endogenous and exogenous)
3. Positioning of SBUs in portfolio
4. Deriving strategies

5. Creating a target portfolio

To apply the procedure to data curation, there are really just two adjustments to make. First, it is essential to know what SBUs mean. Put simply, in Strategic Management SBUs are products or services that define the core competence of the enterprise (cf. Bartol and Martin, 1998). In data curation management this would more closely correspond to the data curation processes and workflows themselves. Second, one has to think about the dimensions. Again, in the most prominent examples of Strategic Management the dimensions are specified as market growth rate (exogenous) and relative market share (endogenous), attractiveness (exogenous) and competitive capabilities (endogenous), or competitive position and product evolution respectively.

As mentioned above when introducing the strategic positioning concept, there are many possibilities that could be considered in the case of data. Really, the dimensions could also be used to define the classification for standardization and positioning as elaborated in the previous section. Yet, the classification should not be too narrow so that as many curation projects as possible are covered by the same analysis. Following the theoretic suggestion in Step 2, an endogenous dimension is, of course, the degree of difference of the data projects, which scales down to how different the data formats and structures actually are. It is probably the most salient feature that all data curators have to cope with when devising data workflows. A good exogenous candidate is the usage behavior of the data users, which is reflected to a certain degree in the software functionality. To the best of the author's knowledge, sustainable data usage is probably the primary goal in data curation.

Having made these adjustments, the data curator is now able to set up a strategy matrix based on the specific qualities of the available data as advised in Steps 3 and 4 (next section). Step 5 will not be considered here, since it is not the primary aim of this study to give an account of workflow optimization.

Deciding Where to Go: Strategic Positioning

Given the limited financial resources, the main challenge in managing the highly diverse data formats is in choosing an appropriate strategy and designing detailed workflows for the curation process. The strategy developed here centers around a subtle analysis of the components of each curation process, where similar data correlate with similar curation workflows. Yet, different data formats usually always require a different curation workflow. In addition, all workflows deviate in the degree of the demands of use often laid down by the scientists themselves or implicitly given by the functionality of the software processing the data. These two variables usage (most often software functionality) and data format are chosen to be binary here for reasons of simplicity. So each variable has two properties, i.e. same and different, that delineate a two dimensional space, in which four strategic fields of data curation workflows can be placed (see Figure 1). The four strategies unfold by logic of the combination of the variable's properties, same and different.

The first strategic field, defined by handling the same data and the same usage, specifies a standard configuration of the repository framework with little to no adjustments. The classic example across all disciplines are lexica, biographies, and glossaries of all sorts or data that have the same structure, such as a bidirectional 1:1-relation. Their usage requirements are searching for entries, adding and deleting entries, and attaching files (text and picture). There are ready-made methods that take care of the import and export of data, input forms and output styles.

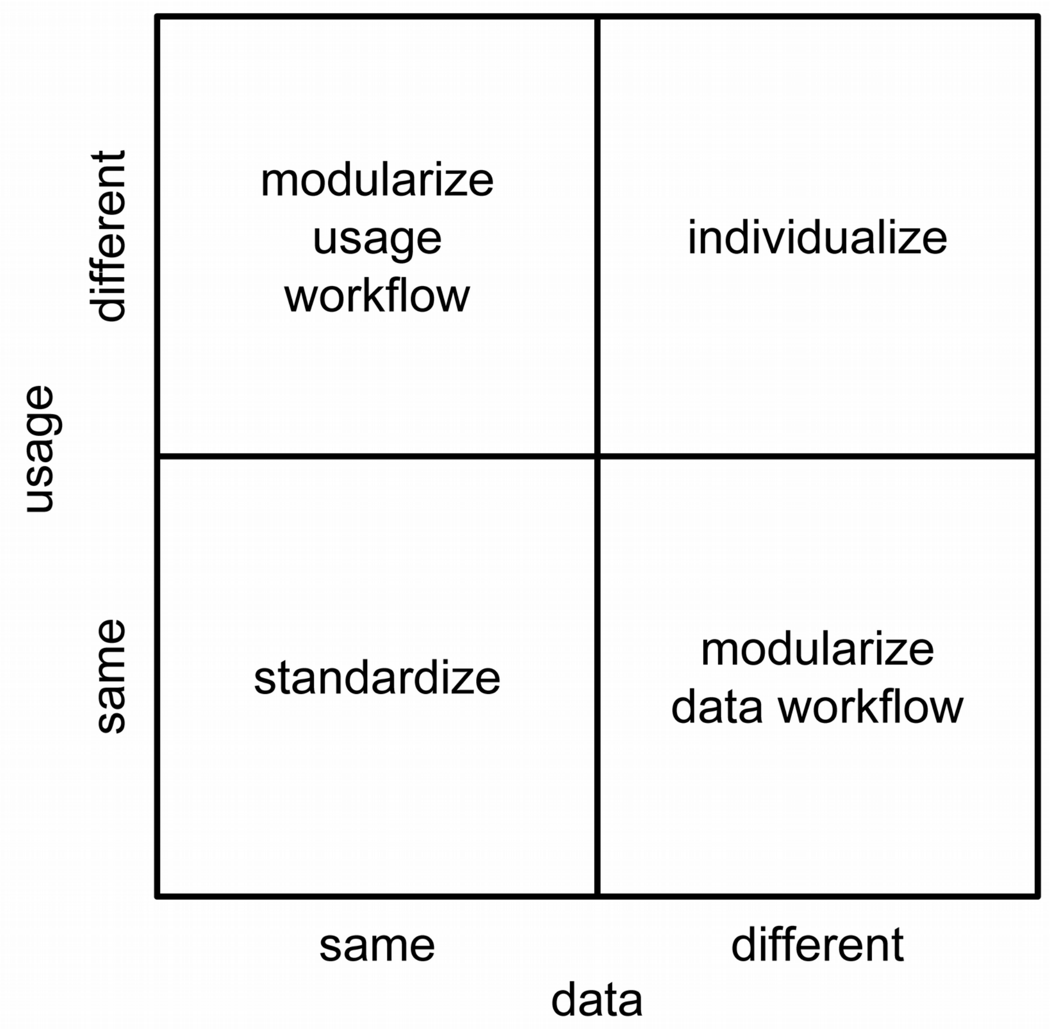

Figure 1. Four-field-strategy matrix.

The second field, a repository specification (same usage – different data), needs more substantial adjustments in designing and implementing new data models. This might also involve the development of metadata schemes and their implementation, however, the software functionality for accessing and manipulating the data remains the same, e.g. search, add, and delete a data set. The usual candidates that this strategy applies to are more complex collections whose structure cannot be scaled down to simple 1:1 relations. To be precise, there are corpora whose sentences and words are modeled with the TEI-standard or historical collections that harbor real subset relations in the representation of dates, places, events, references and sources that partially depend on one another.

The third strategic field specifies a repository, in which data represented by one model, is exploited in several ways. Typically, these comprise survey data from which different software components extract, consolidate, or process data, e.g. for the use of differentiated statistics for various research questions. A different functionality might also allow users to continuously enlarge the stock of questionnaires from different samples. A more specific example are representations of speech signals with consistent metadata schemes. Again, different software applications are embedded to either analyze features from the speech signal or use parts of the annotated data as templates serving as a component for further processing in speech recognition.

Last, repositories deviating in both usage and data formats show the least degree of standardization. They are represented by the fourth strategic field. Most of the time these workflows are designed from scratch because of their idiosyncrasies regarding their interaction of data and application. Yet, they might become part of field two or three if future curation projects show similar data models or software requirements so that the once individualized workflows become part of the standard methodological repertoire.

Enabling Flexibility: Devising Modular Components

The strategic alignment helps to coordinate the future curation project by evaluating its complexity and by setting the boundaries, in which the project's specific workflow can be designed. Nevertheless, the strategy does not tell the data curator anything about the particular processes to choose and their order. To do that, a second classification of the properties of the curation process is needed. An analysis of all curation projects carried out so far resulted in a classification of modules from which concrete tasks can be derived. The present classification is based on a survey carried out in the Faculty of Humanities at the University of Hamburg (Wörner, 2015), but it is open to steady refinement and enlargement if new data formats and applications are encountered that could not be considered so far.

The purpose of the classification is to define modules sharing roughly the same properties of data or usage respectively. These modules can be combined to flexible workflows depending on their strategic positioning in the above matrix. So far, four main types of modules that could be further subdivided in smaller units are identified:

1. data analysis,
2. software analysis,
3. design new metadata schema including new data model, and
4. integrate supplementary software components.

These workflow modules are really the strategic propositions derived from the positioning of the collected curation projects so far, as claimed in Step 3, in any portfolio analysis. The first two modules are always applied. They belong to the standard workflow and serve to position the curation project in the strategy matrix as illustrated in Figure 1. Further steps depend on where the project is positioned. If the positioning results in the quarter named standardize, the standard workflow is used to deposit the data in a ready-made repository with available search and data entry functionality that does not need any further adjustments. Duration frames and re-curation can be estimated with high exactness. The individualize strategy in Figure 1 means to start an independent software project, for which both data models and new software components are developed. There is no standard repository that can be used here and little experience in the duration of finalizing such projects can be given. The workflow assimilates the stages usually involved in software projects (use case, design and architecture, implementation, and testing).

Modularizing the usage workflow proposes to maintain the standard data model and metadata schema, but devise specialized sub-workflows to meet deviant expectations of the users. These sub-workflows depend on the specification of the software functionality. Likewise data workflow modularization suggests working out specialized workflows to define the new data model specific to the project and leave the standard software specification for using the data unchanged. Indeed, it is possible to use the same portfolio technique as described above over again in a new subproblem. And each satisfiably solved set of sub-problems produces modules of sub-workflows that can be exploited in similar data curation projects. In fact, this recursive process is an enhancement of the classic portfolio analysis, which is advantageous in data curation to reach the desired level of detail. The iterative process comes to a halt if the desired level of granularity is obtained. It is important to note that all modules can be used in the individualize-strategy as well.

Discussion

The strategic instruments worked out in the previous sections aim towards a preferably exact evaluation and planning of the curation workflow, the specification of workflow modules show how best to achieve the positioning of a new curation project. The final result of each workflow is a data repository meeting the specific needs of data and usage. As implied in the matrix, there are four possible repository specifications. Each of the four strategies' workflows leads to a different repository configuration. This procedures reads soundly in theory, yet it reveals some misconceptions in practice.

A common criticism of portfolio analysis is that generalizations derived from the matrix are misleading (cf. Bartol and Martin, 1998). More specifically, portfolios consolidate in part very complex interrelations to two dimensions. For the case of data curation this simplification could lead to designing workflows in terms of the data structure-data usage divide, while neglecting e.g. practicability, flexibility, or even acceptability and preferability issues by the data curator. Depending on the emerged actual practices at a data center, the most efficient workflows cannot be identified in the strategy matrix.

A second argument could go along the lines of the concreteness of workflows, that is, the universality of the approach that made it above all applicable to data management also is a major disadvantage. Once again universality is boon and bane at the same time. Strategic thinking usually takes a bird's eye perspective by abstracting away from individual cases. However, workflows in data curation seem to emerge bottom-up more often than top-down. In other words, very specific process knowledge known as best practices is crucial for successful data curation. Looking at data curation phenomena from too much of a distance is severely restricted by missing the specifics that make a difference for the overall strategic positioning. Put more provocatively, the question is how does strategic thinking help data scientists with their daily work of devising workflows that are concrete enough to be directly applied?

Finally, a third argument could be raised on achieving standardization. Workflow and strategy positioning in the portfolio matrix as such are ineffective if no further actions follow from the analysis. Without question, one can see from the matrix what the situation is, but it does not explicitly provide the right measures of action explicating how it should be. There is no indicator showing whether the present state as is reached an optimum simply by latently improving and adjusting workflows based on the experience of the data curators. There are no cues if further standardization is at all possible.

While these disadvantages cannot be denied, it is also clear that the strategic management dimension in data curation workflows will not deliver a holistic solution. In fact, it gives an additional perspective in organizing and planning data workflows from which may (but must not necessarily) follow saving and learning effects. Indeed, strategic portfolios will not explicate detailed workflows, but they help to organize them and make decisions that are more likely to be effective. One can think of the portfolio technique as well-defined corridors, in which further decisions have to made by the data scientist. And yet, these corridors can be narrowed further to a certain degree by extending the portfolios to the more specific situation of each data center. Two examples of how this can be done are shown next.

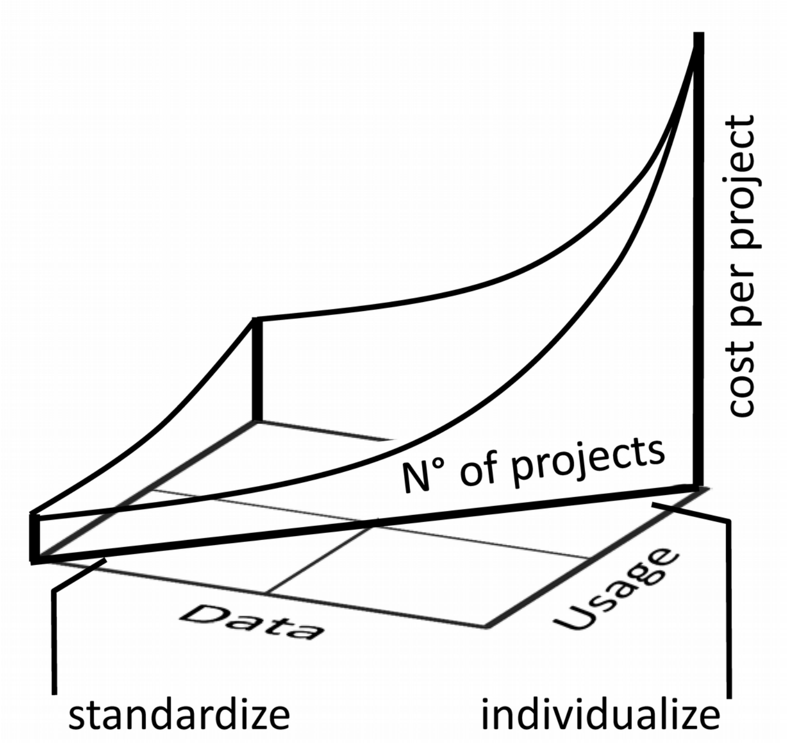

Figure 2. Portfolio analysis and experience curve.

An advantage of portfolio analysis in general is its easy use and extensibility. As argued above, experience curve and lifecycle analysis are strategic instruments that can easily be integrated in the portfolio analysis. To exemplify this, two extensions to the four-field-strategy matrix are shortly introduced here.

In Figure 2 the traditional experience curve concept, in which the production output is mapped to unit costs, is applied to the strategy matrix. The conclusion to be drawn from Figure 2 is not so much that costs of individualized workflows raise exponentially. This is common knowledge. Rather, costs decrease if the number of similar projects increases so that the degree of standardization rises. Pushing this chain of ideas further, it is not the number of curation projects that needs to be similar, but the number, distribution, and partition of involved sub-workflows. As a consequence, even in a world of very heterogenous curation projects, it is possible to achieve high degrees of standardization if the projects can be attributed to a set of sub-workflows that are engaged at high, on average about, equal frequency.

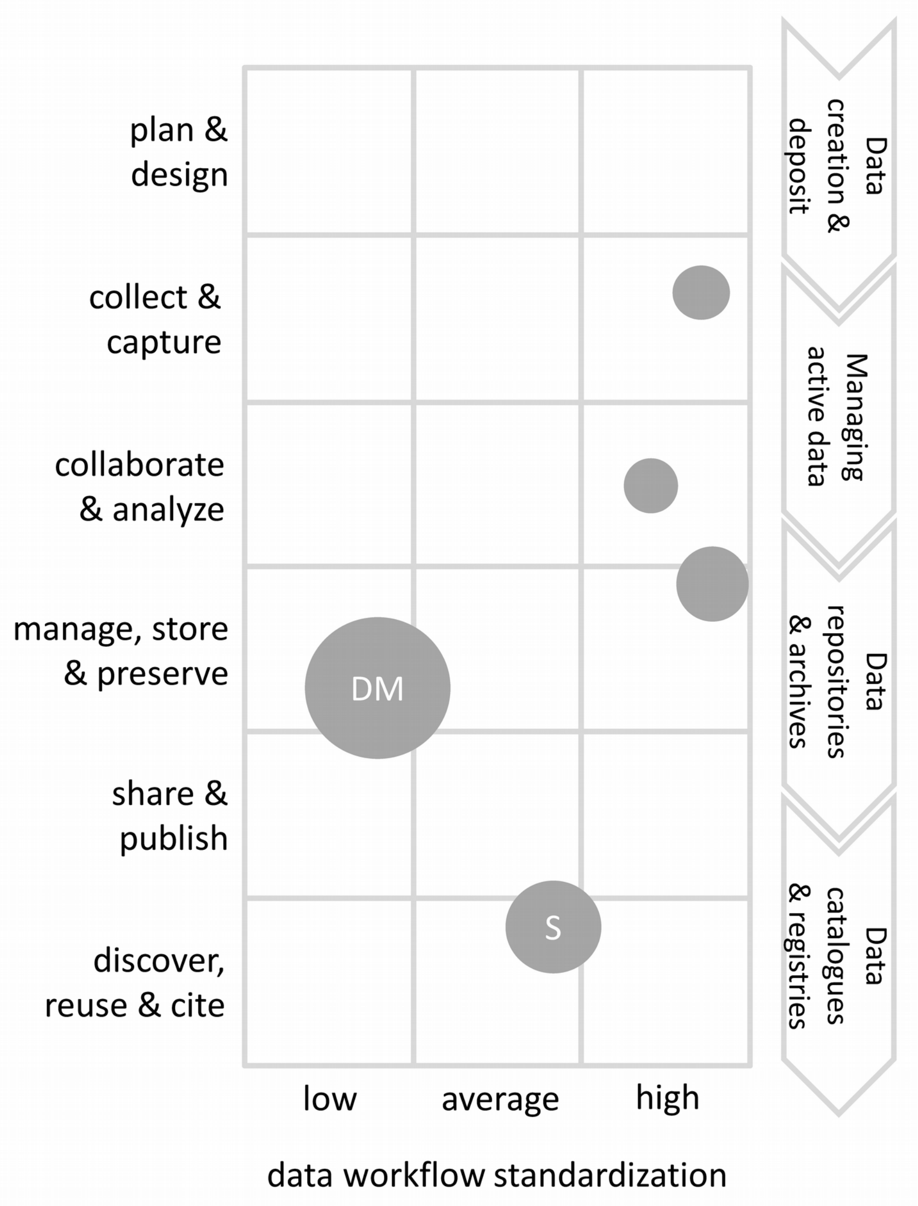

Figure 3. Portfolio analysis and Data Lifecycle

One of the most influential portfolio matrices in Strategic Management is the product evolution matrix (e.g. Bartol and Martin, 1998). This portfolio maps the products competitive position to the product lifecycle. By giving up the endogenous – exogenous dichotomy one can construct an equally expedient matrix for research data. Figure 3 replicates the research data lifecycle[1] on the perpendicular and delineates the degree of standardization on the horizontal axis. Given that data curation is not meant to

[1] Meeting the requirements of the EPSRC research data policy: https://www.jisc.ac.uk/guides/meeting-the-requirements-of-the-EPSRC-research-data-policy

be a subprocess of data archiving,[2] but admits a broader interpretation, that is, data curation is applicable to the entire data lifecycle (Rümpel, 2011), approved workflows can be placed as circles in the matrix whereas the radius of each circle indicates the number of projects, in which the workflow is used (see Figure 3). As an illustration, the DM-workflow (data modelling) is positioned in the matrix. If this is done for all workflows identified in the projects of the data center, the data manager gets a good idea on the distribution of curation workflows at the various stages in the data lifecycle and its share of standardization. Many small placements mean that the projects are very heterogenous. In this case, it is likely that sub-workflow partitioning to a sufficient degree failed and, if workflows were not already restructured, it hints at a potential of standardization effects.

Summary

In summary, much about the approach introduced here is to give a hand on possible ways of designing effective data curation workflows for curating very heterogeneous data while considering usage patterns. Shortly, data curation workflows can be placed somewhere between a continuum of standardizing and individualizing workflows.

When fine tuning the process analysis, i.e. to find out which subprocesses in curating data can exactly be standardized, it becomes clear that even quite idiosyncratic projects reveal intertwined subprocesses, for which standardized solutions exist, and subprocesses, for which no solutions are available. So the problem is the identification and entanglement of data curation workflows into its parts with a subsequent rearrangement and modularization. Portfolio analysis as a strategic instrument, adapted from Strategic Management, takes much of the burden of organizing this task and preserves a clear perspective on the relevant processes. Portfolio analyses can also be used to optimize already established workflows.

Optimization in data management is about increasing the amount of standardization since the experience curve effect is predominant. By positioning curation projects in the strategy matrix, data curators have a firm point of reference of how to identify and rearrange the project and possibly to revert to already well understood processes. Thus, without explicit process management, new pareto optima are accomplished.

Finally, at the current state of research and experimenting with data curation, it deserves mention that finding the right strategy needs time, constant feedback, and steady improvement. It seems to be a permanent process of balancing between the standardization objective and fulfilling the expectations of data users. Thus a state at which the cost of the characteristic long tail of research data is at acceptable levels is approached rather sinuously.

References

Ansoff, H.I. (1965). *Strategic management.* New York: McGraw-Hill.

Bartol, K.M. & Martin, D.C. (1998). *Management.* Boston: McGraw-Hill.

Boston Consulting Group. (1973). *BCG-Perspective.* Boston: BCG. Retrieved from http://www.bcg.de/bcg_deutschland/geschichte/klassiker/portfoliomatrix.aspx

[2] See DARIAH-DE: https://de.dariah.eu/research-data-lifecycle

Dahringer, L.D. & Mühlbacher, H. (1991). *International marketing: A global perspective.* Reading: Cengage Learning Emea.

Ghemawat, P. (2000). Competition and business strategy in historical perspective [Online Paper]. Harvard University – Strategy Unit: HBS Comp. Strategy Working Paper No. 798010. doi:10.2139/ssrn.264528

Hünerberg, R. (1994). *Internationales Marketing.* Landsberg/Lech: verlag moderne industrie.

Jauch, L.R. & Glueck, W.F. (1988). *Business strategy and strategic management.* New York: McGraw-Hill.

Keegan, W.J. (1989). *Global marketing management.* New Jersey: Pearson.

Kotler, P. & Armstrong, G. (2016). *Principles of marketing.* Boston: Pearson.

Kreutzer, R. (1989). *Global marketing – Konzeption eines länderübergreifenden Marketing.* Wiesbaden: Deutscher Universitäts-Verlag.

Nag, R., Hambrick, D.C. & Chen, M.-J. (2007). What is strategic management, really? inductive derivation of a consensus definition of the field. *Strategic Management Journal, 28*, 935-955.

Porter, M. (1980). *Competitive strategy.* New York: Free Press.

Rümpel, S. (2011). Der Lebenszyklus von Forschungsdaten. In S. Büttner & M. L. Hobohm Hans-Christoph (Eds.), *Handbuch forschungsdatenmanagement* (p. 25-34).

Segler, K. (1986). *Basisstrategien im internationalen Marketing.* New York: Campus Verlag GmbH.

Sun, W. (2007). *The art of war: Sun Zi's military methods.* New York: Columbia University Press.

Thompson, A.A. & Strickland, A.J. (1992). *Strategic management: Concept and cases.* Irwin: Homewood.

von Clausewitz, C. (2011). *Strategie: aus dem Jahr 1804, mit Zusätzen von 1808 und 1809*. Zürich: Paradeplatz-Verlag.

von Clausewitz, C. (2012). *Vom Kriege.* Hamburg: Nikol.

Wörner, K. (2015). *Auswertung der Professorenumfrage zum Konzept eHumanities 2020+.* Survey carried out in the Humanities' Department at the University of Hamburg in 2013. Retrieved from http://hdl.handle.net/11858/00-248C-0000-002C-80BB-6