Culture Data Space: A Case Study in Federated Data Ecosystems

Matthias Jarke[1,2,*,†]

[1]RWTH Aachen University, Databases and Information Systems (Computer Science 5), Ahornstrasse 55, 52074 Aachen, Germany
[2]Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven, 53757 Sankt Augustin, Germany

Abstract

In several national and even continental data strategies worldwide, the decentralized data space concept aims to address concerns of data sovereignty among organizations. A significant number of projects and a few already operational data spaces address application domains in industrial domains such as manufacturing, mobility and logistics, or health. However, data sovereignty has also become a key concern of artists and cultural institutions who are pursuing the two-pronged and sometimes conflicting goals of creating added value from data sharing, and protecting their intellectual property and personal privacy rights. Additional challenges of this sector include an orders-of-magnitude larger number of potential players compared to existing data spaces, frequently limited IT capabilities, a complex differentiated system of data types and regulations, and a novel interplay between heterogeneous data integration and analytics with human creativity, among many others. Also, the different evolution paths of the involved sub-communities require a sophisticated concept for federated data space evolution. This keynote talk reports experiences of the "Data Space Culture", a lighthouse project of the German Chancellors Office aiming at investigating these issues and demonstrating and evaluating a suitable data ecosystem around four use cases in the fields of theaters, museums, music training, and networking local culture communities. We also discuss the potential synergies and interoperation challenges with the many other culture digitization initiatives in Europe and beyond.

Keywords

data space, GAIA-X, culture Informatics, data sovereignty, data exchange, data ecosystem, federated data integration

1. Introduction

The European cultural sector is one of the most important fields, characterized by a wide diversity in digitalization. Some large individual players (e.g. the Deutsche Museum in Munich) have invested intensely in all kinds of digital upgrading of their artefacts, or (like the movie and music industries, or the Europeana cultural heritage sector) are organized by large-scale multimedia data collection and streaming platforms, or multi-year joint efforts such as the Europeana or the German Digital Library.

In contrast, many others struggle even with minimal IT infrastructure and limited access to their intended audiences. Moreover, the potential advantages through joint value creation through controlled B2B data sharing without giving up sovereignty of their own data rarely expoited.

The European Data Strategy therefore mentions culture as one important application and innovation domain for the concept of data spaces, but other areas such as logistics, mobility, and manufacturing or even health have been earlier in starting specific initiatives.

Encouraged by early regional initiatives in eCulture, the Culture Department of the German Chancellor's office (BKM Bund) has therefore funded a large-scale experimental effort to set up a Cultural Data Space. The project is led by the German Academy of Science and Technology acatech, the Fraunhofer Institute for Applied Information Technology FIT, and Hamburgs Ministry of Culture and Media. In addition, the project (2022-2025) involves representatives of dozens of cultural organizations and research institutions the culture fields of museums, theaters, music, and local culture information networks.

The analysis of these four use case areas revealed the need for some important extensions of the emerging data space technologies, but also some important implications for governance and business models. In this keynote paper, we give a short summary of this initiative and its linkages to other data space application domains.

2. Background: Data Spaces for Data Sovereignty

The idea of data spaces started in Microsoft research in the early 2000s [1], aiming at a personalized organization for the increasingly heterogeneous swamp of data on personal computers. A major technical challenge was semantic modeling and querying across the different technical storage formats, bringing database technologies

Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW'23) — Data Ecosystems (DEco), August 28 - September 1, 2023, Vancouver, Canada
*Corresponding author.
jarke@dbis.rwth-aachen.de (M. Jarke)
https://www.dbis.rwth-aachen.de/ (M. Jarke)
0000-0001-6169-2942 (M. Jarke)

together with information retrieval as well as semantic networks and later linked open data.

In the 2010s, this original small-scale concept grew into large muli-player systems in two dimensions. First, structured data warehouses of the late 1990s evolved to Data Lakes which store raw data from many subfields or sub-organizations, and allow data cleaning, detection of related data, and sophistcated data analytics and machine learning from many different perspectives, without big upfront integration investment [2].

The second dimension emphasized decentralization of data ownership and data usage control. Driven by the importance of niche players and hidden industry champions in Europe, an intense debate on data sovereignty emerged, e.g. in the GAIA-X initiative [3] which led to European legislations such as the European Data and AI Acts, and the GDPR.

A minimal data space system involves the following core functionalities [4**?** , 5]:

- standardized connectors as a kind of wrapper gateway for the import and export of data to/from the organizational information system of a participant; thus, connectors can serve the technical roles of data supplier, data requester, or both; the units of data exchange are often digital shadows of objects or activities which carry some value [6]
- broker functionality involving one or more catalogs and associated vocabulary support to help the searching and matching between information offers and requests [7];
- contract patterns, contracting workflows, and contract execution monitoring for ensuring sovereign data exchange processes, including the definition and partially automated monitoring of access constraints and usage policies;
- services for the authentification of participants, and the certification of all above-mentioned system components according to rules of the IS association, i.e. protection against misuse of the data space by outsiders

To summarize, the data space technology enables the formation and operation of communities of data exchange and cooperative, secure value creation [8]. A data space is thus complementary to large-scale repositories or data lakes, even though storage can be an elective functionality offered or used by certain players within a data space community.

3. Use Case Requirements for a Culture Data Space

Intense political and technical discussions resulted in the decision to strengthen and validate attractiveness and feasibility of a Culture Data Space by

- exploring the possibility of reusing structures and software components of a somewhat similar already operational data space, in the field of human mobility; not surprisingly, this turned out to be nontrivial, as this data space and even its the underlying base software was and is rapidly evolving.
- identifying the adaption and extension needs of a broad range of cultural fields with specific high-visibility use cases, such that we could address the common core functionalities;
- coping with the complex multi-organizational setting which on the one hand involves cultural institutions and creativity industries at multiple scales, and on the other hand regional interests ranging from local communities to state and federal to even European level.

The four use cases were designed to analyse and showcase specific opportunities and challenges for important individual improvements while jointly generating an understanding of the technical and organizational needs for a Culture Workspace. In the sequel, we briefly summarize the ideas, approach, and challenges for these use cases.

In the last few years, a large number of *city- or county-wide culture platforms* have emerged independently from each other, including many creative ideas and tools, such as, for example, linking information about cultural events to local public transportation systems, thus creating synergy between more cultural interest, touristic value, and better environmental sustainability. The use case aims firstly at enabling interoperability between such cultural platforms (e.g. for creating tour proposals across regional boundaries), but also with large-scale data repositories such as the German Digital Library, the more than 1.500 regional archives, and even media companies. The overload created by this enormously extended offerings will be reduced by culture-specific personalization tools across this heterogeneous landscape, without resorting to a centralization of the data. This use case is jointly coordinated by the OWL culture platform and the IS department of Paderborn University.

Museum exhibitions often face the challenge that only a few original works related to the theme – each often worth millions – are available locally, others must be borrowed and insured at enormous cost, or in digital form. The reuse of artefacts and processes for later purposes

tends to be extremely limited. The second use case aims to demonstrate how the Cultural Data Space could reduce these problems in the context of a sequence of cooperating on a series of exhibitions on the occasion of the 250th anniversary of the pioneering painter of the romantic period, Caspar David Friedrich. Each of the main exhibitions in Hamburg, Berlin, Dresden, and New York will interpret his works and related complementary materials in a different way, i.e. in data management speak, we would talk about multiple highly complex views on overlapping sets of information across many locations and organizational settings. Core research challenges here include federated information collection and presentation, but also very sophisticated legal processes and inter-organizational workflows with high security demands. Coordinating partner for this use case is Kunsthalle Hamburg.

Currently, *theaters* create and disseminate their play schedules and programs mostly in an extremely labor-intensive ad-hoc fashion. Especially short-term changes often require repetition of the whole process, or disappointed audiences. The aim of this use case is to significantly simplify these processes, and even their interoperation, through semantic data standards and associated software tools that leave enough room for individual creativity in advertising, yet facilitate incremental change, interoperation and joint offerings with others. Core partner of this use case is the German Buehnenverein which represents over forty different organizations of theaters (including e.g. the Theater of Augsburg) and similar cultural institutions and their employees and related free artists.

Amateur music training and performing is a confusing market between millions of amateur musicians and students, and many thousands of teachers and conductors. Improving the matching between students and teachers is a key challenge not just for many individuals, but also for small to medium church and laymens choirs. Moreover, once such groupings have been formed, they also want to play music jointly, sharing background information and enjoying latency-free presentation across different locations. Technologically, the music market place this use case is aiming at thus combines features of domain-specific dating platforms with real-time multimedia conferencing, again considering copyright and other usage control aspects as well as the assignment of the created value. While basic system components for this use case exist for selected high-end settings, their main challenge is the scalability to very high number of participants, whereas current data space applications typically involve only a few dozen to hundreds of data space members. The core partner here is the Conservatory of Hamburg, a music teaching with strong digitization and international music teaching experience.

4. Technological and Organizational Implications

Summarizing the results of the use case analyses, we can identify the following general characteristics of a Culture Data Space:

First, culture requires a clear differentiation of important kinds of data which underlie different regulations as well as data management and analytics. At the core are 'works of art' including their digital shadows [9] which may have high value and are subject to ownership, lineage [10], copyrights, and other regulations (e.g. export controls). A second important category are usage and transactional data streams subject to CRM, visitor statistics, and other important data mining tasks. A third group, overlapping with the former, are personal data subject to the GDPR regulations as well as specific personalization methods. Last not least, there is an extremely rich set of possible metadata, following different standards or practice in each cultural subcommunity.

Related to this basic differentiation, at least the following requirements result:

- a scaling of participant numbers over previously studied data space applications by at least two orders of magnitude;
- a culture-specific extension of access and usage policies, including the embedding of existing monetization organizations such as, in Germany, VG Wort, VG Media, or GEMA for musical performance rights;
- to avoid double work and unnecessary inconsistencies, the creation of interoperability and functional synergy with large public or private data collections, such as the German and European digital libraries„ public archives, and media organizations, but also compatibilities with the onging efforts towards domain-specific FAIR-compatible research data infrastructures [11], such as NFDI4Culture and also the European Data Space Support Center;
- methods and tools for semantic interoperability among the many existing and forthcoming metadata standards in the various cultural domains but also to related data spaces such as Mobility, Tourism, European Cultural Heritage, and the like.
- in addition to the above-mentioned control-level and metadata infrastructure, also culture-adapted optimization for the actual data exchange and value-added processing, including data-kind specific storage, query,integration, and personalization services in the highly heterogeneous setting, following a logic-based approach as in [12].

These in part unique requirements imply careful thinking about a suitable data space architecture. Among the over 20 open and private proposed infrasstructures on the market, we considered three open source variants for the Culture Data Space architecture.

The early data space applications, such as the German Mobility Data Space, bundle all the core data space functions – except the connectors of the participants in a single operational organization which provided the core services such as broker, contract, and identity management. It seemed natural to start with such an already operational infrastructure for the first experiments, and define a growth path for the new requirements.

For data spaces with tech-savvy large industrial participants, such as the emerging catena-X data space for the automotive industry, this architecture was considered to limited, but also too intrusive since e.g. usage policies were intended to be enforced into the individual connectors of participants, such that they felt to lose some control over their own information systems. They preferred a solution in which they take full responsibility for linking their internal system to connectors. The connectors themselves have only a very basic functionality, and the participants can add those of the formerly central services they need as so-called extensions to the connectors. The open source development program for this effort is mostly performed in the EDC connector initiative of the Eclipse foundation by industrial and science partners including Fraunhofer.

However, this approach assumes high IT-/data management competence of at least the largest part of data space participants, which is totally unrealistic for most of the cultural sector. Moreover, different user groups and sub-dataspace communities progress at very different speeds, e.g. depending on whether they can invest their own money or need to rely on (initial) public support. Even from the experiences within the current project, it seems unimaginable to think of a single governing body which could make sufficiently quick and mutually acceptable decisions. We therefore opted for a third architecture which consists of several federated sub-dataspaces run by a culture domain, a very large data repository organization, a city or rural community, or perhaps by a creative industry organisation which would offer similar services as the original IDS approach but e.g. federated brokers and contract support for data exchange across the boundaries of the subcommunity.

5. Conclusion and Outlook

Our case studies have shown remarkable additional attractions and requirements for a Culture Data Space. We stress again that our understanding of such a data space focuses on the sovereign sharing of different kinds of data, and the value creation and fair value appropriation [13]. It is thus complementary to the large data collection efforts and, abstractly spoken, to the data lake approach.

At present, an initial "sandbox" data space has been set up at Fraunhofer which includes first extensions to the central approach, in particular aiming at federated brokers with intelligent metadata translation facilities. By the end of 2023, two of the four use cases will be in part operational, all of them fully a year later. In parallel, the generic topics of advanced rights management, the governance structure of the federated data space ecosystem, and the future operating companies, and the business models will be decided.

Acknowledgments

This work was supported in part by the Beauftragte fuer Kultur und Medien der Bundesregierung as a lighthouse project Datenraum Kultur within the digialization strategy of the German Federal Government. Special thanks go to Georgios Toubekis who coordinated the complex use case analyses.

References

[1] A. Halevy, M. Franklin, D. Maier, Principles of data space systems, in: 25th ACM PODS, ACM Press, 2006, pp. 751–772.

[2] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: a survey of functions and systems, IEEE Trans. Knowledge and Data Engineering 35 (2023) 1–20.

[3] W. G. GAIA-X, Reference Architecture Document, Release 21.03, GAIA-X, European Association for Data and Cloud AISBL, 2021.

[4] M. Jarke, C. Quix, On warehouses, lakes, and spaces – the changing role of conceptual modeling for data integration, in: J. Cabot et al. (eds.): Conceptual Modeling Perspectives, SpringerNature, 2017, pp. 231–245.

[5] B. Otto, M. ten Hompel, S. Wrobel, Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, SpringerNature, 2022.

[6] M. Jarke, Data sovereignty and the internet of production, in: International Conference on Advanced Information Systems Engineering – CAiSE 20, Springer, 2020, pp. 549–558.

[7] R. Fagin, P. G. Kolaitis, R. J. Miller, L. Popa, Data exchange: semantics and query answering, Theoretical Computer Science 336 (2005) 89–124.

[8] S. Geisler, M.-E. Vidal, C. Cappiello, B. Bernadette Farias Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, J. Rehof, Knowledge-driven data ecosystems towards data

transparency, ACM Journal of Data and Information Quality JDIQ 14 (2022) 1–13.

[9] M. Liebenberg, M. Jarke, Information Systems Engineering with Digital Shadows: Concept and Case Studies, in: International Conference on Advanced Information Systems Engineering – CAiSE 20, Springer, 2020, pp. 70–84.

[10] Y. Cui, J. Widom, Lineage tracing for general data warehouse transformations, VLDB Journal 12 (2003) 41–58.

[11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific data 3 (2016).

[12] M. Lenzerini, Direct and Reverse Rewriting in Data Interoperability, in: International Conference on Advanced Information Systems Engineering, Springer, 2019, pp. 3–13.

[13] F. Piller, V. Nitsch, D. Luettgens, A. Mertens, S. Puetz, M. van Dyck, Forecasting Next Generation Manufacturing: Digital Shadows, Human-Machine Collaboration, and Data-Driven Business Models, Springer Contributions to Management Science, 2022.