IJDC | *Conference Paper*

Transparent Disclosure, Curation & Preservation of Dynamic Digital Resources

Deirdre Lungley
UK Data Service, UK Data Archive

Darren Bell
UK Data Service, UK Data Archive

Hervé L'Hours
UK Data Service, UK Data Archive

Abstract

This paper explores an enhanced curation lifecycle being developed at the UK Data Service (UKDS), with our Data Product Builder. Through a Graphical User Interface, we aim to provide the researcher with a tailored digital resource. We detail the threefold motivation behind this initiative: data dissemination scalability, researcher satisfaction and the reduction of nationwide duplication of research effort.

Subsequent sections detail the technical components and challenges involved. In addition to more standard data subsetting, filtering and linking components, this data dissemination platform offers dynamic disclosure assessments – identifying combinations of variables that present a potential disclosure risk. All components are underpinned by the Data Documentation Initiative's new Cross-Domain Integration standard (DDI-CDI), designed to handle the many structures in which data may be organised.

Ever conscious of the scale of the task we are embarking on, we remain motivated by the need for such advances in data dissemination and optimistic of the feasibility of such a system to meet the needs of the researcher while balancing the data disclosivity concerns of the data depositor.

Submitted 9 February 2024 ~ *Accepted* 22 February 2024

Correspondence should be addressed to Deirdre Lungley, UK Data Archive, University of Essex, Wivenhoe, Colchester, Essex, CO4 3SQ. Email: dmlung@essex.ac.uk

This paper was presented at the International Digital Curation Conference IDCC24, 19-21 February 2024

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.v18i1.937
DOI: 10.2218/ijdc.v18i1.937

Introduction

The traditional and potentially overused metaphor of the research data lifecycle has some significant attractions and benefits. It highlights the benefits of cyclical science where outputs are openly shared, ideally through deposit in a repository that sets deposit standards, undertakes appraisal, applies appropriate levels of curation, and potentially offers long term preservation to ensure that digital objects remain usable and understandable over time. The repository provides discovery and access services, and the results of subsequent research are themselves re-deposited in a virtuous circle.

But presentation of the lifecycle as a basic sequence of deposit, curate/preserve, discover and access can hide some of the complex activities and functions of (meta)data services, and oversimplify the approach to individual digital objects. The lifecycle model may also tend towards assumptions that a single, concrete digital object containing traditional file formats is being conveyed through workflow steps by a monolithic organisational entity. In reality repositories and related data and metadata services are often delivered through complex and dynamic partnerships of organisations and systems. Moreover, the expectations and assumptions about traditional digital objects are being increasingly challenged by the emergence of linked open (meta)data.

In repositories where some or all of the data and metadata being curated, accessed and used is sensitive in some way, the challenges of managing legal, ethical and information security issues increase. The societal benefits of research into sensitive data, particularly that which risks disclosure of information about human subjects, must be balanced against the trust that data subjects, researchers, repositories and funders have in the measures taken to ensure data is appropriately protected and used. It is often the case for traditional file-driven digital objects that an overall disclosure risk evaluation about a dataset, results in the assignment of a sensitivity classification that in turn drives the choice of access terms and methods, in line with the Five Safes framework (Ritchie, 2017, Arbuckle & Ritchie, 2019). In addition to safe data held in appropriate environments that require evaluation and approval or (safe) outputs, this model defines a need for appropriately approved research projects undertaken by trained and accredited researchers (safe people), within 'safe settings', i.e., the assignment of a higher risk factor may imply that all of the files that make up the digital object can only be accessed within a controlled environment such as a safe room or via secure remote access (a safe setting).

This bibliographic assignment of a sensitivity classification drawn from a controlled vocabulary of options can be augmented by a more granular data engineering approach that enables the underlying data sources to be expressed as dynamic digital resources, rather than static documentary artefacts. At the UK Data Service (UKDS) we are exploring an enhanced curation lifecycle with our Data Product Builder (DPB). Through a Graphical User Interface, we aim to provide the researcher with a tailored digital resource that will meet their particular research need. Output variables (columns) can be selected, respondents (rows) can be filtered, datasets can be linked dynamically, and users are offered a choice of disclosure mitigations to address the risk which typically emerge from dataset linkage. With this method we can progress beyond static datasets labelled with one of our three sensitivity classification levels: 'open', 'safeguarded', 'secure'. Through dynamic and interactive feedback on the variable level disclosure implications of each subset or linkage choice, the researcher will have the opportunity to craft a derived digital object where the sensitivity classification is communicated in real time.

The motivation behind this initiative is threefold: data dissemination scalability, researcher satisfaction and the reduction of nationwide duplication of research effort.

1. An increased awareness of the risks of data disclosure means that data depositors are more likely to seek secure access solutions for their datasets by default. This has the potential to render the entire research lifecycle inoperable as researcher accreditation, project approvals and output checking on increased volumes of data become unsustainable. Can we through rigorous and transparent disclosure risk computation

provide a mechanism for disseminating more data safely outside these secure environments?

2. To increase researcher satisfaction, can we support dynamically built datasets which are customised to their analytical requirements instead of monolithic digital objects with a single sensitivity classification? For example, a particular researcher studying the employment prospects of young people may consider detailed age to be essential but is prepared to compromise on detailed occupation/industry classifications or we may find that the confined age window chosen allows both detailed age and such classifications.

3. We are also increasingly conscious of the duplication of effort occurring in research facilities all over the UK, where multiple individuals and teams are developing similar research resources, for example linking datasets using geographical ontologies or scraping the web for various classification systems. By selecting, curating, preserving and presenting the researcher with one of our prebuilt dataset linkages and providing our resulting data product with rich metadata including ontological labels we aim to reduce the workload of the researcher.

Researchers that can meet the various project approval, training and technical criteria necessary to access sensitive data will seek to extract relevant subsets of data (with potentially differing sensitivity levels from the overall source data) and link these with other subsets. Linked subsets in turn have potentially different levels of disclosure risk from their source data. Currently permission to extract data output from a secure environment, e.g. to provide supporting data for a published paper, depends on analogue, human-mediated assessment of their disclosure risk. Our DPB aims to transparently and dynamically compute sensitivity and give the researcher iterative opportunities to modify their selection, including top/bottom coding and re-banding, to allow them immediate access to data that still meets their research needs.

Methodology

Providing dynamic digital resources through our Data Product Builder, requires both comprehensive metadata and powerful computation. Together they enable dynamic data selection, linking, filtering and output serialisation. These derived data products are candidates for appraisal, retention and long-term preservation based on their value for reuse and as the evidential basis of publications. This platform in addition employs dynamic disclosure risk analysis —identifying combinations of variables that present a potential disclosure risk— and offers the researcher suitable mitigations.

Metadata Enrichment

The DPB is underpinned by the Data Documentation Initiative’s Cross-Domain Integration standard (DDI-CDI)[1]. DDI-CDI is designed to handle the many structures in which data may be organised, including ‘wide’ data (rectangular) and ‘multi-dimensional’ (cross-tabulated) and it manages meaning through concepts, allowing these concepts to perform different roles in different structures. These features support navigation between the study data, held in a DDI-CDI Wide Data Set, and the related Census aggregate data, held in a DDI-CDI Dimensional Data Set. Additionally, our classification ontologies, e.g., the occupational classification systems we use, are stored as SKOS[2] concept systems. Figure 1 illustrates the flow of metadata and data through our DPB prototype. The DDI-CDI and SKOS data feed an interactive user interface, allowing the researcher to build their custom data product, which in turn can be available as a

[1] https://ddialliance.org/Specification/ddi-cdi

[2] https://www.w3.org/2004/02/skos/

dataset within the system. We are also generating these custom datasets as RDF triples and importing them into a Fuseki[3] instance. An authenticated API over this triple store, powered by SPARQL[4] queries, allows us to explore the potential of linked data.

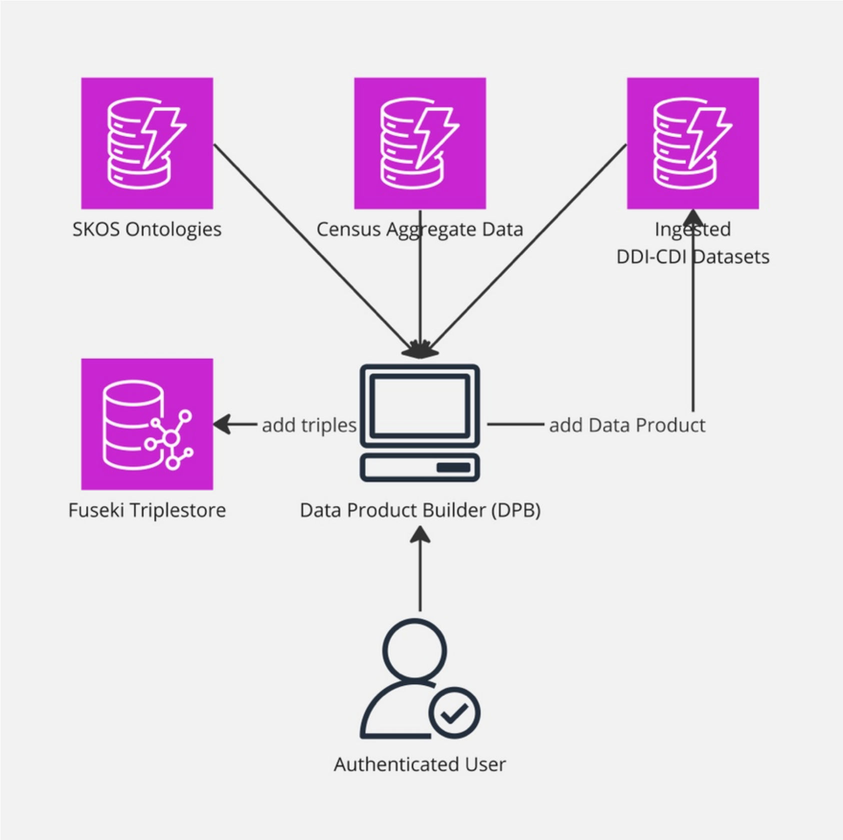


Figure 1. UKDS Data Product Builder meta(data) flow.

The choice of DDI-CDI has been driven to some extent by its inherent interoperability but also by the potential for us to take advantage of its provenance structures. DDI-CDI would allow us to provide detailed, machine-actionable metadata at the level of a data point – a DDI-CDI datum. This would enable the maintenance of provenance information –the sources and processes that were used to create each data point, from the original dataset to the resultant data product. In future iterations of the DPB we envisage using DDI-CDI's Process model to capture further details of provenance.

An ingest pipeline orchestrates the metadata enrichment processes needed to transform our existing archival metadata, currently stored as DDI-Codebook (DDI-C), to the formats detailed above which are necessary for an "enhanced publishing platform". The UKDS is in the process of moving to the more powerful DDI-Lifecycle (DDI-L) metadata standard, for our more traditional data dissemination platform used to make binary datasets available to researchers. These DDI-L assets will be stored in a Colectica Repository[5] and once this is fully implemented, a future DPB iteration will use this repository's API in the initial step of its ingest pipeline. Traditionally, datasets from the social science domain tend to store microdata in formats such as SPSS or STATA and the ingest pipeline we are developing currently combines DDI-C metadata with SPSS microdata to create the DDI-CDI datasets, driving the DPB. However, increasingly repositories are seeing demand for data from a wider range of domains, which implies the need to handle a wider range of formats, e.g. environmental datasets stored in

[3] https://jena.apache.org/documentation/fuseki2/

[4] https://www.w3.org/TR/sparql11-overview/

[5] https://www.colectica.com/software/repository/

compact grid formats, and therefore we are exploring the ingest of such data using DDI-CDI's 'long' data format, designed for event data.

Machine Learning

Ingesting DDI-CDI datasets, data and detailed metadata, requires successive enrichment steps to elevate it from the relatively limited DDI-C metadata and the raw binary SPSS format. Driving an enhanced dissemination platform with the CDI format requires such elements as variable level semantic concepts and sensitivity concepts, and Census harmonised variable representations. Scaling this metadata enrichment to thousands of datasets cannot be done manually but cannot yet be undertaken in a fully automated way. We are currently training machine learning models to feed an active learning cycle, one which allows manual intervention and augments its learning data with this new input. An example of such learning is the identification of key variables – the use of machine learning classification models to annotate variables with Data Privacy Vocabulary (DPV)[6] concepts – a critical step in the curation process. We are using the term 'key variables' here as commonly defined in statistical disclosure control (Templ et al., 2015) – a set of variables that, when considered together, can be used to indicate individual units.

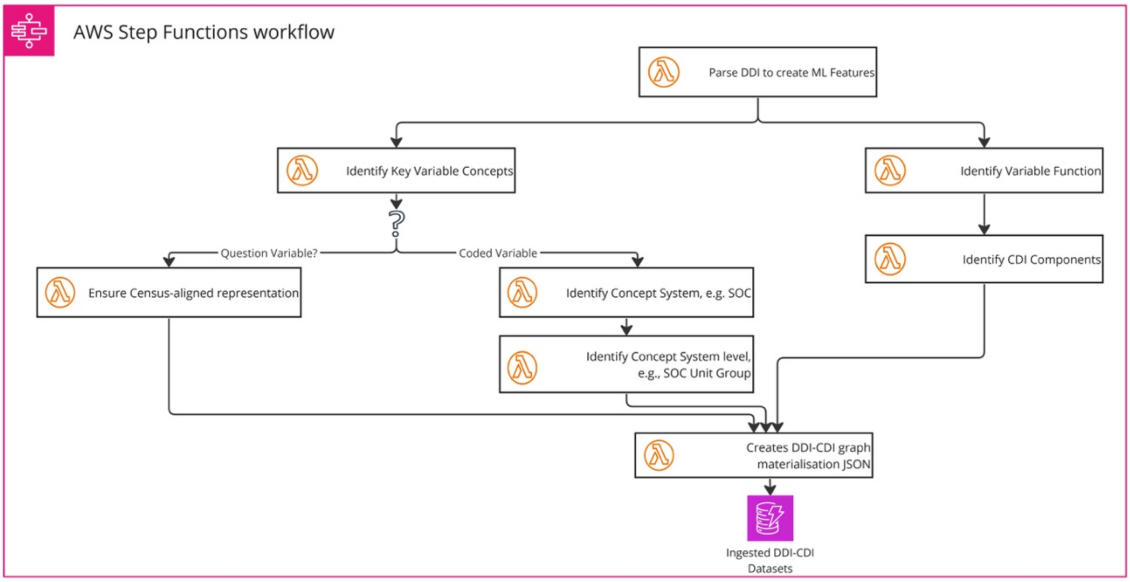

AWS Step Functions workflow
Parse DDI to create ML Features
Identify Key Variable Concepts
Identify Variable Function
?
Question Variable?
Coded Variable
Identify CDI Components
Ensure Census-aligned representation
Identify Concept System, e.g. SOC
Identify Concept System level, e.g., SOC Unit Group
Creates DDI-CDI graph materialisation JSON
Ingested DDI-CDI Datasets

Figure 2. Ingest steps required to create DDI-CDI datasets.

A data dissemination platform offering dynamic disclosure assessments requires detailed variable level metadata. Figure 2 details the classification and harmonisation steps involved. Key variables can be subdivided into question variables and derived/coded variables, requiring different treatments. The latter generally conforms to standard classification hierarchies, an example being the socio-demographic variable, occupation code. For such variables, learning the semantic concept, e.g., Standard Occupation Classification (SOC), is not sufficient, it would need to be further classified as either SOC2010 or SOC2020 and even further to annotate the variable with the level within the hierarchy. Is this SOC Major Group (1 digit), SOC Sub-major Group (2 digits), SOC Minor Group (3 digits) or SOC Unit Group (4 digits)? It is only when we have metadata annotated as such that we can drive disclosure violation checks and mitigations – knowledge of the exact classification hierarchy and the level of the variable within that hierarchy allows us to traverse upwards to a less granular level when required.

[6] https://w3c.github.io/dpv/dpv/

Table 1. The initial subset of key variables.

Key Variables	
Question Variables	**Derived/Coded Variables**
Sex	SOC (Occupations)
Age	SIC (Industries)
Marital status	NS-SeC (Socio-economic)
Economic activity	Geography
Ethnic group	
National identity	
Country of birth	
Language	
Religion	
Health	
Education	

However, with question related variables, e.g., a variable related to a question regarding marital status, the variable representation, the codes and categories of its response domain, may not conform to a particular classification system. Since, as will be detailed in subsequent sections, we are interested in assessing population level disclosure violations using Census aggregate counts, it is necessary for us to classify our key variables as conforming to a classification available in Census data. This in many instances may require harmonisation of the study level representation (codes and categories) to the available Census representation. This is currently the least developed aspect of the DPB prototype, but the one which will be the focus of the coming phase. In collaboration with relevant partner organisations, we are interested in taking advantage of recent advances in pre-trained large language models (LLMs) by fine tuning them on our question and variable representation corpus.

Table 1 lists the initial subset of key variables we are classifying and harmonising. Further desirable classification steps would include identifying the variable function, e.g., is this a weight variable? and the type of CDI component involved, i.e., is this variable an identifier, a measure or an attribute?

Ingest Pipeline

We are iteratively building our data curation ingest pipeline using AWS Step Functions to orchestrate the parsing, classification and harmonising stages necessary to facilitate discovery and automated disclosure control. Its ease of use, clear progress visibility and human intervention capabilities could help increase curator efficiency, in addition to giving these data professionals confidence in the semi-automated ingestion process. After any classification step in the pipeline, it is possible to present the curator with a graphical display of the automated output, e.g., the determined key variables and associated probabilities, with the ability to amend where necessary and feed the learning data.

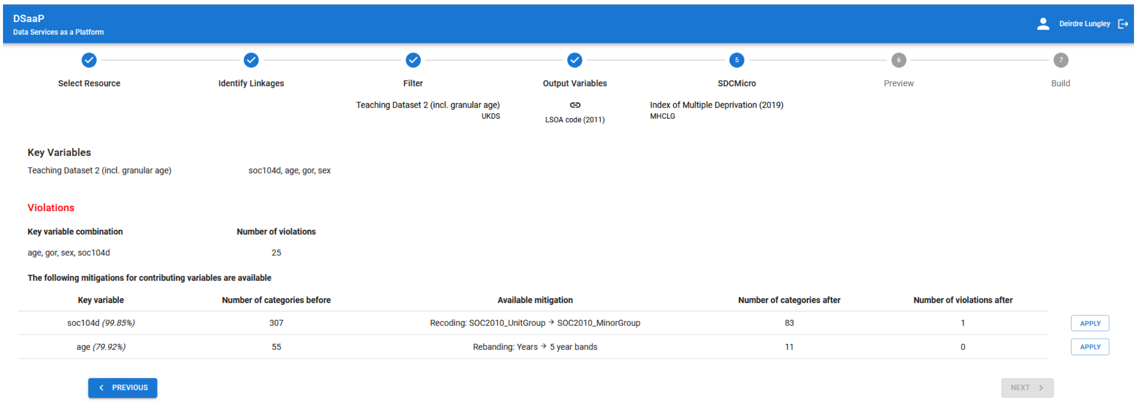
The result of this ingest pipeline is an integrated DDI-CDI corpus with metadata and data combined. This when combined with our SKOS ontologies, containing our classification systems, and the Census aggregate data for the combinations currently available, feed the interactive DPB.

Disclosure Risk Analysis (DRA)

Disclosure risk analysis is performed in the Data Product Builder when the researcher has determined the selected resource, with possible linkages, record and column filters and output variables, i.e., on the resulting 'data product'. Our DRA algorithm is based on sdcMicro (Templ et al., 2015) functions, specifically its frequency calculation algorithm (which underpins its k-Anonymity function) and its SUDA2 function. Frequency calculations are done for all 2-way, 3-way and 4-way combinations of variables which have been classified as "key variables", socio-demographic characteristics such as age, sex and place of birth. Frequency violations occur where the number of instances of a combination of key variables falls below a threshold. This study level threshold is currently set at three. We have then prototyped a further step using Census aggregate data to check population level frequencies for these study level violations. Our prototype is currently limited to publicly available Census aggregate combinations, e.g. age, sex, region, 4-digit occupation code. However, we are exploring extending this to all 4-way combinations of key variables. We treat this population level data as a proxy for the knowledge of the expert curator as we develop the automation of our curation processes. Population level violations are deemed to have occurred when the Census frequency of the variable combination is less than the population level threshold, currently set at five.

The SUDA2 algorithm, which calculates the contribution of each variable to the combined sensitivity outcome, is used to determine which variable mitigations to offer. Currently a mitigation is offered for all variables with a contribution of greater than 50%. Examples of such mitigations would be the re-banding of 4-digit Standard Occupation Classification codes (SOC Unit Group) to the equivalent 3-digit code (SOC Minor Group) or the re-banding of a detailed age variable to 5-year bands. The user can choose the mitigation with the least impact on their research objective. The sub-concept level classification we perform, detailed earlier in this paper, combined with our SKOS ontologies, allow us to traverse any classification hierarchy to a less granular level when combination frequencies determine the data as being too disclosive.

Figure 3 illustrates the power of automated disclosure analysis. In this example the researcher has chosen both a granular occupation variable, SOC Unit Group (4 digits) and a detailed age variable in their output selection. Frequency calculations have determined study level violations and subsequently population level violations. Unless mitigated the researcher would be unable to proceed and download this data. A choice of mitigations is offered, either decreasing the granularity of the occupation variable to SOC Minor Group (3 digits) or re-banding age in 5-year age bands. The latter will remove all violations, the former will require a further mitigation to SOC Sub-major group (2 digits).



Teaching Dataset 2 (incl. granular age)
UKDS

LSOA code (2011)

Index of Multiple Deprivation (2019)
MHCLG

Key Variables

Teaching Dataset 2 (incl. granular age) soc104d, age, gor, sex

Violations

Key variable combination	Number of violations
age, gor, sex, soc104d	25

The following mitigations for contributing variables are available

Key variable	Number of categories before	Available mitigation	Number of categories after	Number of violations after	
soc104d *(99.85%)*	307	Recoding: SOC2010_UnitGroup → SOC2010_MinorGroup	83	1	APPLY
age *(79.92%)*	55	Rebanding: Years → 5 year bands	11	0	APPLY

Figure 3. Data Product Builder disclosure violations and mitigations.

Computational Power

To understand the computational power required for such dynamic DRA and mitigations, we need to examine the scenario where a researcher has chosen a data product which contains six key variables. In fact, this is the limit we have set for key variable selection, to ensure reasonable interactive performance. Such a number of key variables would result in frequency calculations for 15 4-way combinations. If any of these combinations resulted in a frequency threshold violation at sample level, the Census-harmonised values of this variable combination would be checked against Census aggregate combination counts to ensure no violation at population level. If population-level violations do arise, for every violation and for every possible mitigation, new frequency calculations are performed on the mitigated copies of the data and the researcher is presented with the resulting options to reduce disclosivity. The researcher can iterate many times over the data, adding and retracting choices as necessary, and each time these calculations are rerun until the resulting data product is no longer deemed disclosive. In order to make this level of computation feasible in real-time, we have re-engineered the frequency calculation algorithm to avail of C++ bitmask operations in place of the original R code. In addition, cloud-based hardware allows us to dynamically scale computational power as required.

With regard to real-time discovery, our ingested DDI-CDI datasets are manifested as an extremely large NoSQL cloud database, optimised for sub-millisecond querying. Combined with auxiliary cloud computational resources, it enables real-time combinatorial experiments by the researcher, allowing genuinely data-driven (rather than keyword-driven) discovery – interactively the researcher is informed of exactly how many records meet their selection.

Next Steps

We have already detailed a number of improvements to this prototype which have either been initiated or are planned for the coming months, e.g., improvements in data provenance utilising DDI-CDI's Datum and Process model. Here we detail other potential enhancements for future iterations.

In addition to the rapid, statistically rigorous and transparent risk assessments the system currently offers, we could offer adjusted parameters per dataset, per researcher/research project and per variable. Thresholds, currently set globally, could after negotiation with each data depositor, be set at a far more granular level, using an appropriate ontology. We are currently exploring the Open Digital Rights Language (ODRL) ontology for this purpose.

We are developing the serialisation options for the data products generated by this system. As mentioned earlier, in addition to the SPSS/Stata/text outputs required by the traditional 'desktop' research world, we are developing RDF triples for linked data environments, by serialising the resulting metadata and data as DDI-CDI triples, validated dynamically by ShEx/SHACL shapes, before being published to a secure Fuseki instance. Currently these triples are exposed by some limited authenticated API endpoints, as we explore the potential of machine actionable linked data. The opportunities of such linked data systems augment and complement file-driven ecosystems rather than replace them. Currently much linked data is derived from statistical data deposits (e.g. in SPSS, STATA or R formats) and with effective curatorial versioning the linked data environment may be used to update and 'round trip' data back into alternative or updated versions of these serialisations. This leveraging of the linked data capabilities can provide source material for traditional research formats and provide efficient and scalable preservation solutions where file format transformations are required.

We are aware that there remain questions on the curation/preservation lifecycle of the outputs of our dynamic dissemination tool. The data products generated by sub-setting and linkage in turn become a part of the curation and preservation ecosystem. Dynamic digital resources increase the complexity of the challenge, but also increase the flexibility of curation, preservation, access and reuse options. Rapid and effective disclosure analysis will allow some of these outputs to be made available through less secure access methods, potentially by other service providers. Each data product derived from subsets and linkage has its own appraisal needs that will guide decisions about whether and for how long it should be retained, and the degree to which it should be further curated and preserved. Snapshots of data that provide evidence for published research may need to be retained 'as is'. Subsets and linkages with the potential for ongoing reuse may need to be preserved and their continued usability augmented, e.g. through updating geographical or occupational ontologies. If the original data resource is not retained, then the data products cannot be dynamically regenerated by re-querying the source.

Challenges

Offering dynamic digital resources in addition to our current static dissemination methods is not without its challenges. In this paper we have detailed many of the technical issues associated with providing such a data dissemination platform and the measures implemented to alleviate them. However, we do not trivialise the specific challenge of population level violation determination using Census aggregate data and addressing this, including the feasibility of storing, accessing and computing such vast volumes of data, is the focus of much of our current development.

Not all of the challenges are technical. It is vitally important to impress on data depositors, data curators and data reusers the potential, even the necessity, of such a transformation of the data lifecycle.

Traditionally, social scientists have been accustomed to downloading datasets to their local machines, exploring the metadata and data available and developing their research hypothesis over time. This is in contrast to other research fields, e.g. health research, where researchers tend to request data downloads post hypothesis construction. We do not wish to hinder such exploratory research but to facilitate it, in fact we are seeking to make this a more dynamic and less time-consuming activity.

Conclusion

At the outset we detailed our threefold motivation for dynamic digital objects: data dissemination scalability, researcher satisfaction and the reduction of nationwide duplication of researcher effort.

Disclosure analysis with associated run time mitigation has the potential to increase trust in sensitivity assessments that lead to research within 'safe settings' and also to simplify and scale valuable research against less sensitive subsets in more open settings. The transparency of the run-time Disclosure Risk Analysis (DRA) algorithms employed, gives data depositors trust in the sensitivity assessments of the data being released.

The potential benefits of dynamic responses to queries that seek to derive subsets of or links between these digital resources go beyond the application of machine-driven empirical disclosure control. These custom choices ensure that the resulting data product has the greatest possible chance of meeting the research needs of the user. In addition, this ability to link, filter and mitigate risk prior to download, frees up researcher time to focus on research rather than data manipulation.

In summary, by harnessing sophisticated metadata and modern computational power we have the opportunity to improve the transparency of the data curation lifecycle and hence boost the trust placed in us by our depositor and researcher community, in addition to better meeting their needs.

Acknowledgements

The authors benefit from UKDS core funding from the ESRC and the University of Essex. We are grateful to our UKDS colleagues, especially Simon George, Thomas Gilders, Ivan Evdokimov, Agnes Bircher and Artiom Rumiancev for their contributions to this project.

References

Arbuckle, L., & Ritchie, F. (2019). The five safes of risk-based anonymization. *IEEE Security and Privacy*, *17(5),* 84-89. Retrieved from https://doi.org/10.1109/MSEC.2019.2929282

Ritchie, F. (2017). The 'Five Safes': a framework for planning, designing and evaluating data access solutions. Paper presented at the *Data for Policy 2017, London, UK*. Retrieved from http://dx.doi.org/10.5281/zenodo.897821

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software, 67(4),* 1-36. Retrieved from https://doi.org/10.18637/jss.v067.i04