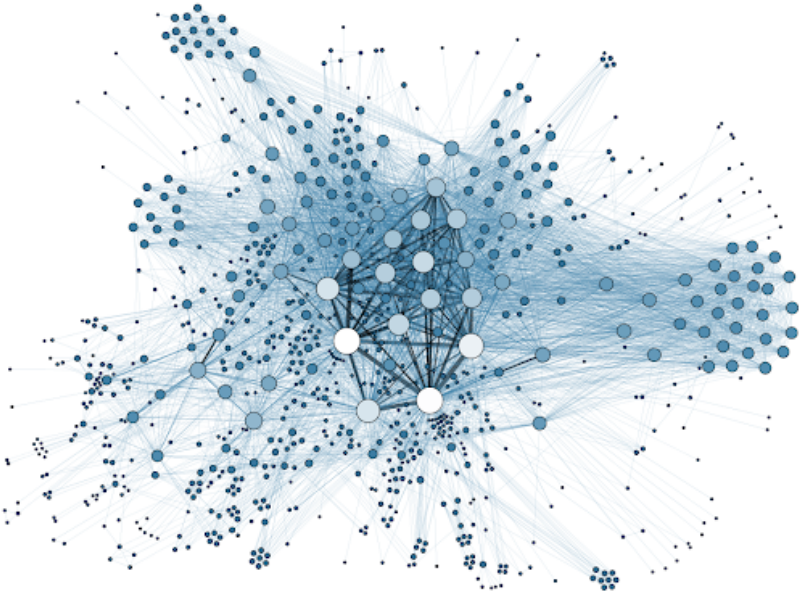
Archaeology Data Service



Linked Data Strategy

Document Control Grid

Author/Contributor	Version	Date of revision	Revision Note	Status
Tim Evans	1.1.	27/08/2020	First Draft	DRAFT
Tim Evans, Holly Wright, Jenny O'Brien	1.2	23/09/2020	Minor edits and corrections	DRAFT
Julian Richards	1.3	28/09/2020	Minor edits	DRAFT
Tim Evans	1.4	02/10/2020	Final edits	LIVE

1. Introduction

Since our beginnings, the Archaeology Data Service (ADS) has endeavoured to align itself with agreed International standards for the provision of Open Data. In later years, this has developed with internal standards-driven workflows incorporating Linked Data within our metadata, and also research-driven attempts to provide ADS (meta)data as Linked Open Data (LOD).

Neither of these two aspects of work have ever been clearly defined within internal policy, and the infrastructure to support internal use of Linked Data, and thereafter the delivery of our (meta)data as LOD has never been a consistent part of ADS Strategy. As the ADS begins to redevelop our organisational Strategies and develop responses to initiatives such as FAIR, it is important to rectify this situation.

This document aims to provide a definitive vision for all aspects of Linked Data within ADS workflows: primarily what we want to achieve, and also a preliminary roadmap for doing so.

Although a stand-alone document, the ADS Linked Data Strategy is part of a number of core documents, and will feed into future iterations of Curatorial Strategy, Metadata Standards, and ADS Annual, Five Year, and Ten Year Strategic Objectives.

2. Background

A short history of Linked Data at the ADS

The ADS has always adhered to metadata standards and vocabularies, initially those required as part of the Arts and Humanities Data Service (AHDS), but evolving into those relevant to the wider UK heritage Sector (principally in England, and using thesauri recommended through MIDAS), and. In all cases, where a controlled terminology was required, then manual entry of the term

alone (“string literals”) were used (either as a drop-down list or free text), as in most cases these were not available in Linked Data formats.

Between 2009 and 2011, the ADS in partnership with the University of South Wales Hypermedia Research Unit and Historic England, undertook a series of small AHRC projects - STAR and STELLAR - to develop new methods for linking archived datasets, unpublished reports, and vocabularies and exploiting the potential of a high level, core ontology and natural language processing techniques. One of the ADS outcomes was the mapping and extraction of archaeological datasets into RDF/XML representation conforming to the CIDOC CRM-EH standard ontology, and thereafter publishing this data as Linked Data via a new triplestore and SPARQL endpoint. While the STAR and STELLAR research objectives were met, there was still a lack of vocabulary control (with unique identifiers) that hindered the full potential of the resulting Linked Data. Although STELLAR tools were capable of generating controlled types, the resulting linked data still currently employ free text strings (including data from the ADS).

Despite the lack of vocabulary control for heritage vocabularies, the success of this project and the implementation of our own triplestore led to an internal drive to push forwards on ADS linked data production. In 2011 an internal Research and Development project by Holly Wright looked to explore the potential of converting ‘traditional’ relational databases held by the ADS, into Linked Open Data. This project used the high-profile Roman Amphora Database as a pilot, and created a SKOSsified thesaurus within the ADS Triplestore. This resource was cited frequently within the associated literature at the time, and used within the data visualization and exploratory interface built in the framework of the ERC Advanced Grant Project (EPNet).

In 2012 an internal development project sought to "triplify" the ADS Collections Management System (CMS) and publish Collections metadata as Linked Data. An internal data structure was established, based on the Dublin Core Metadata Elements. The internal database changes incorporated the first ever use and storage of controlled vocabularies with URIs. This was the Library of Congress Subject Headings (LCSH), which were downloaded as RDF, imported to the

ADS triplestore, and a basic search facility built into the CMS. It is important to note that this was the only successful “triplification” at the time; a separate function mapped and converted literal strings such as ‘language’ and ‘publisher’ to predefined entities in vocabularies including Geonames, Ordnance Survey, and DBpedia. For example ‘Coverage = England’ would always be exported as http://data.ordnancesurvey.co.uk/id/country/england. This prototype created over 1000 Linked Data nodes, and an alternative way of surfacing ADS metadata in a machine readable format.

In 2013, another AHRC project - SENESCHAL - led by the University of South Wales Hypermedia Research Group in collaboration with the ADS and project partners, including Historic England, RCAHMS, and RCAHMW successfully converted the traditional heritage thesauri used within the United Kingdom into Linked Data. The project also created a series of services that enabled ADS to map our historical literal strings (such as ‘Henge’) to the Linked Data vocabularies, thus storing the string and URI in our database. A new module within the CMS was created to use a web service at the newly created heritagedata.org.uk to query these new vocabularies.

Since that point, the only development of note has been work undertaken as part of the ARIADNE project, to develop a European data infrastructure and search portal. ARIADNE uses the Getty Art and Architecture Thesaurus (AAT) for subject terms, and accordingly all UK heritage terms used by the ADS were mapped to the AAT using a methodology and service developed by the ARIADNE project.

A subsequent follow up project, ARIADNEPlus, has improved the previous data model (ACDM) into the ARIADNE Catalogue Ontology (or OA-Cat), an application profile of the CIDOC CRM. ADS is in the process of mapping key datasets, including Archives metadata, to the OA-Cat and importing data into the ARIADNEplus AC triplestore (from which indices for the new Portal are generated) as part of a standard workflow. At the time of writing ARIADNEplus is planning to develop APIs to provide access to the whole AC (including the Catalogue) as Linked Open Data, thereby potentially allowing the ADS to implement their own portal or service in the future.

Current Use of Linked Data within ADS metadata

The ADS Collections metadata currently uses:

- LCSH subject headings (as DC.Subject), currently generated using a search of a locally hosted (ADS triplestore) copy of the LCSH terms, periodically downloaded from the Library of Congress website.
- UK Heritage thesauri (DC.Subject) using web services from Heritage Data. It is important to note that this does **not** include temporal terms.
- Getty Thesaurus of Geographic Names (as DC.Coverage): but **not** stored as true Linked Data (literal string and URI concatenated in a single field).
- ORCID Identifiers: recorded on an ad-hoc basis in a central database, but not consistently stored in Object metadata.

Current provision of ADS (meta)data as LOD

The ADS Linked Open Data are made available as a direct result of the STELLAR project, a joint venture between the University of South Wales, the ADS and Historic England (formerly known as English Heritage).

The ADS Linked Data repository contains triples for each ADS Archive released up to the end of 2015, created as part of an internal pilot project.

The ADS also provides a SPARQL query web interface which can be used to interrogate the ADS tiple store directly.

Current Challenges We Face

1. **INFRASTRUCTURE**: The infrastructure to support use of Linked Data internally has grown organically (as detailed above). For example, only some parts of ADS database systems are set up to support storage of URIs. Certain Interfaces are constructed to work with LOD web services (for querying), whilst others use a variety of methods to incorporate terms on an ad-hoc basis. The current triplestore is maintained on a best-efforts basis, but is now relatively antiquated.

2. **AWARENESS**: Linked Data approaches have traditionally been the domain of a small number of staff, and usually associated with the Research Projects noted above. Although elements such as the Vocabulary URIs are explained to new members of staff, often the rationale ("bigger picture") for the Linked Data approach is often not discussed. This can lead to a knowledge gap when staff leave, and an overall disconnect between curatorial and technical aspects of ADS roles.

3. **EXPERTISE**: The skills and knowledge to construct programmatic functions to generate Linked Data (either as Linked Data formats or through maintenance and upgrading of the ADS triplestore) is not currently optimal within the ADS team.

4. **SCOPE**: Recent work has highlighted two areas of ambiguity regarding ADS use of Linked Data; do we want all our metadata to be Linked Data (i.e. for every field to use LOD vocabularies)?, and which vocabularies and systems do we want/need to use?

3. Vision statement

To ensure our metadata is available as Linked Open Data, and uses vocabularies interoperable with specialist Heritage partners, Funders, and the public sector, and to advocate innovative reuse and collaboration.

4. Mission statement

- Optimise ADS **Consumption** of LOD vocabularies: To develop our internal systems and policies to increase the capability to record our collections and object metadata using Linked Data vocabularies.
- Investigate efficient and appropriate **Production** of Linked Open Data**:** To establish if ADS can undertake this work, and if so the best method and workflow for ensuring (meta)data is available as LOD.

Overarching Goal

To have use of Linked Data vocabularies as an embedded part of our Curatorial and Preservation workflows, and to develop partnerships and methods that allow us to have our (meta) available as Linked Open Data, but within the capabilities of a small organisation.

5. Strategic objectives

Understand our Metadata (LDS 1)

1.1. Review all metadata used for Collections and Objects, and produce a clear definition of which data should/shouldn't use a LOD vocabulary or Persistent Digital Identifier. This includes a review against recommendations from partners such as ARIADNE, DataCite, FISH, NERC, and MEDIN.

1.2. Define what vocabularies or schemes this data should use. This should include a review of the current status of:

- Dbpedia+WikiData;
- British Library BNB;
- TNA PRONOM;
- GeoNames;
- Ordnance Survey;
- Getty Vocabularies;
- NERC;
- ISO 639-2 Codes or Lexvo;
- Virtual International Authority File (VIAF);

1.3. Produce a '**Linked Data Standards**' document for internal use that records this decision, and which directly feeds into the ADS Cataloguing Policy.

1.4. Inform, and Implement a consistent approach to (linked) metadata capture across all ingest platforms including ADS-EASY, Library, and OASIS.

1.5. Establish an annual review of use of Linked Data within our metadata, attended by all Staff.

Improve Internal Recording (LDS 2)

2.1. Feed into Curatorial Objectives to ensure that CMS redevelopments facilitate the recording of Linked Data vocabularies and Persistent Digital Identifiers for all data fields required.

2.2. Review and Establish *how* we Consume Linked Data, factoring in risk of use of external web services, and cost of local hosting.

- Establish which data can, and should, be consumed via use of web services.
- Establish which data can be hosted locally, and develop procedures for ensuring local versions are updated on a regular basis.
- Establish procedure for manual entry of term and URI.
- Establish policy and procedure for adding new vocabularies or schemes.

2.3. Develop a clear specification for CMS redevelopment.

- Ensure that the capacity to add new thesauri is included.

Decommission Current Triplestore (LDS 3)

3.1. Create an RDF download from the current ADS tripestore.

3.2. Make RDF available to download from the ADS Metadata Page.

3.3. Establish the best method of replacing LCSH triples with a localised solution.

- Feed this into any redevelopments required as part of 2.3

3.4. Decommission Triplestore and Update any existing documentation.

Develop a Methodology for Publishing Collection Metadata available as Linked Open Data (LDS 4)

4.1. Assess capability for an internal methodology (i.e. export and conversion to RDF from Oracle) for generation of LOD.

4.2. Work with partners from ARIADNE to establish possible external workflows/Services for generating LOD.

4.3. Establish the best option, dependencies, cost, and timeframe for the new LOD delivery method.

Develop a Recommendation for Publishing Object Metadata available as Linked Open Data (LDS 5)

- Assess capability for internal methodology (i.e. export and conversion to RDF from Oracle) for generation of LOD.
- Work with partners from ARIADNE to establish possible external workflows/Services for generating LOD from granular data.
- Establish the best option, dependencies, cost, and methodology (including selection criteria) for the new LOD delivery method.

Increase Awareness within ADS and Stay Up-to-Date with trends in Linked Data (LDS 6)

6.1. Schedule lunchtime Seminar introducing this document and the history of the ADS and Linked Data.

6.2. Incorporate the Linked Data Strategy into the quarterly reviews of the Curatorial Objectives to ensure communication between infrastructure development, Cataloguing Policy, and Research and Development.

6.3. Publish a paper, or Conference Paper on ADS and LOD.

6. Roadmap

The following Roadmap will be updated every year, pending progress against key objectives and the results of Reviews and Partnerships

Year 1: 2020/2021

- Schedule lunchtime Seminar introducing this document and the history of the ADS and Linked Data (6.1). To be completed **by January 2021**.
- Undertake review of CMS and OMS (1.1 and 1.2), to be completed **by January 2021**

- Create Standards document and that this feeds back into all relevant documentation and Policy (1.3), to be completed **by March 2021**
- Review and Establish how we Consume Linked Data, factoring in risk of use of external web services, and cost of local hosting. Use this to write a short report on future work that is required (2.1 and 2.2), **by August 2021**.
- Assess capability for internal methodology for Publishing LOD (4.1 and 5.1). A simple appraisal and decision to be made by Systems Manager and Deputy Director **by August 2021**.
- Decommission current Triplestore (3.1-3.4). To be completed **by April 2021**.
- Ensure regular quarterly meetings between key staff (Deputy Director, Curatorial Lead, International Projects Manager, Director) to review progress and to inform Strategy and Roadmap for collaboration with ARIADNE.