RO-Index: A survey of Research Object usage

This manuscript (permalink) was automatically generated from stain/ro-index-paper@c9bb0fb on August 20, 2019.

Authors

- **Stian Soiland-Reyes**
  0000-0001-9842-9718 · stain · soilandreyes
  Department of Computer Science, The University of Manchester, UK; Informatics Institute, Faculty of Science, University of Amsterdam, NL · Funded by BioExcel-2 (European Commission H2020-INFRAEDI-02-2018-823830)

- **Paul Groth**
  0000-0003-0183-6910 · pgroth
  Informatics Institute, Faculty of Science, University of Amsterdam, NL

Abstract

> This manuscript is **work in progress** and (for now) follows the style of a Study Protocol for F1000Research Registered Reports

For this study we aim to build **RO-Index**, a broad and comprehensive corpus of Research Objects found "in the wild". The proposed methodology follows multiple strands to find the "breeding grounds" of research objects and further describes how Research Objects are selected for inclusion, along with post-processing to build the corpus.

The corpus of Research Objects will primarily be distributed as Open Data, including:

- Identifiers and access URLs
- Extracted manifests and annotation files
- Checksums/references for external data
- Metadata from repository (e.g. Datacite XML)
- Provenance of data gathering and post-processing

Research Objects that cannot be redistributed (e.g. unknown license) will only be examined for aggregates.

A brief set of qualitative and quantitative analytics will then be performed across the overall corpus, in particular to address research questions like:

- Where are Research Objects published?
- Which scientific domains produce Research Objects?
- What serializations are used for making Research Objects?
- What vocabularies are used to describe Research Objects and their content?
- What type of resources to Research Objects contain?
- What kinds of life cycles do Research Objects follow?

Introduction

Protocol

Finding Research Objects

One goal of this work is to determine what kind of artifacts, in practice, can be considered a *research object*. For the purpose of building a corpus we need to have both inclusion and exclusion criteria.

The foundational article on the RO concept is [1] and its workshop predecessor [2]. The Research Object community has maintained lists of initiatives and Research Object profiles which provide curated, although potentially biased, collections of Research Object approaches and implementations.

Declared Research Object usage

In order to determine potential sources of Research Objects we will start with these community lists, but expand based on a literature review by following any academic citation of the before-mentioned Research Object articles to find potential repositories, tools and communities that may conceptually claim to have or make "research objects". This is a broad interpretation that does not expand into general datasets or packaging formats. The list may be expanded by literate search for "Research Object", the RO vocabularies and standard URLs.

Each of the citing articles will then be assessed to see if they have openly accessible research objects that are possible to identify, and ideally retrieve, by building a programmatic crawler. Ideally such access would use an open harvesting protocol like OAI-PMH or ResourceSync, but it is predicted that in the majority of cases custom crawler code will need to be developed per repository, in addition to manual harvesting of identifiers for smaller collections and individual Research Objects.

Keyword searches

In addition to this "self-claimed" research object usage we will search in more general repositories by developing a list of keywords like "research object", "robundle" or the RO vocabulary URLs. We will search in at least:

- https://github.com/
- https://gitlab.com/
- http://datacite.org/
- https://zenodo.org/
- https://toolbox.google.com/datasetsearch/
- https://dataverse.harvard.edu/

It is predicted that these searches will yield duplicates, but will be used to find potentially new Research Object sources or free-standing instances.

Archives with manifests

Finally we will consider broadly Open Data repositories of file archives (e.g. ZIP, tar.gz) to inspect for the presence of a *manifest*-like file (e.g. `/manifest.rdf`). For practical reasons this search will be restricted to a smaller selection of public repositories and formats, e.g. Zenodo (20k *.zip Datasets), FigShare ("zip" Datasets), Mendeley Data "zip" File Set.

A list of trigger filename patterns will be developed, including:

- `META-INF/manifest.xml` and `META-INF/container.xml` from EPUB Open Container Format
- `manifest.xml` from COMBINE archives [3]
- `.ro/manifest.rdf` from RO Hub [4]
- `.ro/manifest.json` from Research Object Bundle [5]
- `metadata/manifest.json` from RO-Bagit and BDBag [6;]
- `CATALOG.json` from DataCrate [7]

- `ro-crate-metadata.jsonld` from RO-Crate [8]

It is predicted that most of the archive files will *not* contain such a manifest, therefore they can be inspected "on the fly" by the crawler without intermediate storage, to first detect a short-list of archives that contain a manifest-like file. These can then be downloaded in full for further inspection. File-name matching will inspect potential sub-directories, e.g. to detect `nested/data/manifest.xml`, but will classify these archives differently from direct matches.

Candidate sources

For each candidate source we will collect and assess:

- Date assessed
- Assessed by
- URL
- Name
- Estimate # ROs
- Estimate # users
- Maintainer/publisher
- Community links (if any)
- RO profile/format (if any)
- Identifier scheme(s) (if any)
- Persistence/Versioning (if any)

Then for each candidate source we will evaluate:

- Accessability - can we retrieve RO and/or their metadata
- License - permissions and/or restrictions to redistribute the ROs and/or their metadata
- Feasibility - can we programmatically retrieve all ROs (or just a sample)?
- Duplication - could the "same" RO be present by multiple identifiers or in other repositories?
- Self-identified - are Research Objects classified as such (or using similar terminology)?

We may contact the provider or maintainer to expand on these questions if unclear from public information, however we are not conducting a formal survey, as our main interest lays in the machine-readable information from the research objects themselves.

We will finally form a shortlist of sources for further harvesting, considering:

- Programmatically access (or interesting enough to warrant manual access)
- Diversity - might this source be different from the majority of sources?
- Legality - are we allowed to retrieve ROs (or their identifiers and metadata?)
- Confidentiality - are the research objects accessible to the public? (anonymous access or access by 'fresh' user registration)

Handling personally identifiable information

Research Objects may, by their nature, contain information about people and their research activities. It is therefore important that our data collection, processing and potential re-distribution is in consistent with the General Data Protection Regulation (GDPR). To this end we will evaluate:

- Does the source have a GDPR-compliant privacy policy or equivalent?
- Is personally identifiable information contained by identifier (e.g. username)?
- May personally identifiable information be contained by the Research Object manifest/description

- May personally identifiable information be contained by the Research Object files/content?
- Does the RO (or the metadata) have a license that permits redistribution and attribution, e.g. Creative Commons Attribution 4.0 (CC-BY)?

Evaluating this may require retrieving research objects in the first place, but particular care will be taken to classify Research Objects and their sources according to the above evaluation in order to filter information that can progress to be part of the Open Data RO-Index corpus. This forms a staged inclusion list:

1. Unfiltered list of identifiers for a source will be shared if the identifiers tend not to include personally identifiable information
2. Metadata will be shared if it is accessible and does not tend to include personally identifiable information
3. Metadata and identifier will be shared if an open attribution-permitting license is indicated (or implied by site)
4. Content/files will be shared if accessible and an open license is indicated (or, for archives, implied by archive license)

Note: In the above, "tend to" will be determined manually by inspecting a smaller subset of typically 10 research objects. The selection will aim to approximate a simple random subset, but may need to be expanded to take into account the overall diversity of ROs at the source, e.g. date, authors, subsystem, formats. The identifiers of the ROs of this subset will be recorded, along with a description of how the subset was selected.

The inclusion list may be further restricted based on findings from further processing (e.g. a repository is found to distribute sensitive data).

It is worth noting that compliance with open licenses like Creative Commons Attribution 4.0 (CC-BY) or Apache License 2.0 **require** attribution to be propagated (if present). Attribution may sometimes take the form of a URL, identifier, project or organization which do not directly identify a person.

The inclusion list will form different subsets of Research Objects:

1. Identified Research Objects
2. "Non-sensitive" (but potentially closed) metadata
3. Open metadata (potentially personally identifiable)
4. Open content (potentially personally identifiable)

Data for any excluded Research Objects will only be kept for the purpose and duration of this study on computer infrastructure managed by The University of Manchester. Data from excluded Research Objects will only be used for non-person-identifiable aggregated results (e.g. number of CSV files) and broad categorization (e.g. vocabularies used in metadata).

The identifiers from category 1, metadata from category 3 and data from category 4 will be shared in the public Open Data repository Zenodo according to Zenodo's policies. Metadata from category 3 and 4 above may be exposed for programmatic querying (e.g. SPARQL) or converted to other formats. No additional linking with internal and external data sources will be performed, although the collected Research Objects may already contain such links (e.g. https://orcid.org/ identifiers of authors); an exception to this rule is that linking will be permitted to detect duplicate Research Objects across multiple sources, and to access resources clearly *aggregated* as part of the Research Object.

For GDPR purposes the *Data Controller* is The University of Manchester, data subjects may contact `info@esciencelab.org.uk` for any enquiries, such as to request access to data about themselves,

or to request update or removal of personally identifiable information.

Pre-identified data sources

Proto-research objects

- myExperiment packs
- COMBINE archives [3,9]
- VoID datasets http://www.openphacts.org/specs/2013/WD-datadesc-20130912/ [10,3]
- DataONE Data packages [11]

ORE-based research objects

- CWL Viewer https://view.commonwl.org/workflows [12]
- RO Bundle https://w3id.org/bundle/2014-11-05/ [5]
- Workflow PROV corpus [13]
- CWLProv 10.1093/gigascience/giz095 aka [14]
- http://www.rohub.org/ [4]
- http://rohub.linkeddata.es/
- SEEK: https://fairdomhub.org/investigations
- BDBags with MinID [6;]
- Zenodo e.g. [15]
- Mendeley Data eg [16]
- Maven https://repository.mygrid.org.uk/artifactory/ops/org/openphacts/data/
- DocumentObject https://github.com/binfalse/DocumentObjectCompiler/
- GitHub search
- EOSC-Life (too early?)

Software/container-based research objects

- https://sci-f.github.io/ [17]
- https://frictionlessdata.io/specs/data-package/

2nd generation ROs

- DataCrate: https://github.com/UTS-eResearch/datacrate/blob/master/spec/1.0/data_crate_specification_v1.0.md#examples
- RO-Crate: https://data.research.uts.edu.au/examples/ro-crate/0.2/

Manifest formats

A key characteristic of a Research Object is the presence of a *manifest* that describes and relates the content. However, multiple potential formats and conventions have emerged for how to serialize such a format. (..)

Proposed workflow

The overall data gathering workflow is envisioned as:

1. Traverse repository (or one of its sub-sections) using API like OAI-PMH
2. Filter for entries that have an archive-like file type (e.g. ZIP, tar.gz)
3. Retrieve entry's Datacite-like metadata from repository (e.g. DOI, author, license)
4. Start downloading archive
5. Stream archive though a utility like sunzip to list filenames within
6. Record filenames mapped to identifier

7. Select entries which have a manifest-like file in list
8. Re-download selected archives
9. Extract manifest(s) from archives
10. Classify manifests based on format and vocabulary (e.g. RDF/XML using ORE-OAI)
11. Record provenance of data gathering

Post-processing workflow:

1. Convert manifests to a unified RDF format (e.g. N-Triples)
2. Populate quad store (e.g. Apache Jena) with converted manifests
3.

https://zenodo.org/communities/ro/?page=1&size=20

https://developers.zenodo.org/#metadata-formats

Prototype workflow

A prototype workflow is being developed using Common Workflow Language [18], figure 1 shows how the content of a ZIP file can be listed in a streaming mode.

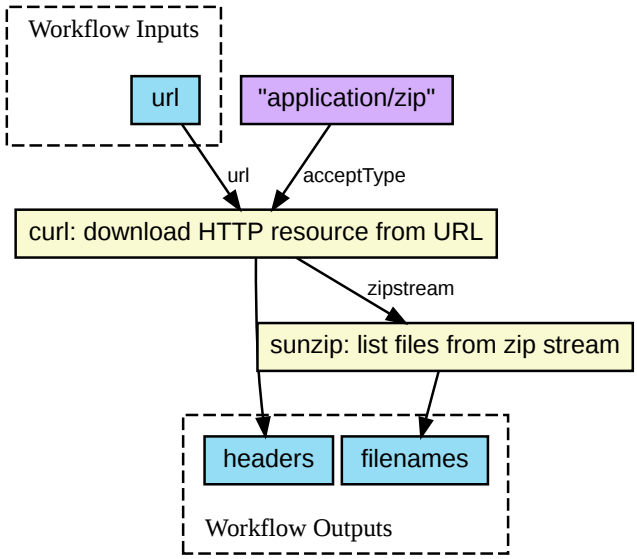

Figure 1: **CWL workflow: List ZIP content by URL** Visualization by CWL Viewer https://w3id.org/cwl/view/git/cd01d30ffc9e04b8804b62df5e985ebfa6f5b276/code/data-gathering/workflows/zip-content-by-url.cwl

Conclusions/Discussion

Data (and Software) Availability

Author contributions

- Conceptualization:
- Data Curation:
- Formal Analysis:
- Funding Acquisition: SSR, CAG

- Investigation:
- Methodology:
- Project Administration:
- Resources: CAG, SSR
- Software: SSR
- Supervision: PG
- Validation:
- Visualization:
- Writing – Original Draft Preparation: SSR
- Writing – Review & Editing:

Competing interests

Grant Information

This work has been done as part of the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union contracts H2020-INFRAEDI-02-2018-823830, H2020-EINFRA-2015-1-675728.

References

1. **Why linked data is not enough for scientists**
Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, … Carole Goble
Future Generation Computer Systems (2013-02) https://doi.org/bgmqrb
DOI: 10.1016/j.future.2011.08.004

2. **Why Linked Data is Not Enough for Scientists**
Sean Bechhofer, John Ainsworth, Jiten Bhagat, Iain Buchan, Philip Couch, Don Cruickshank, David De Roure, Mark Delderfield, Ian Dunlop, Matthew Gamble, … Shoaib Sufi
2010 IEEE Sixth International Conference on e-Science (2010-12) https://doi.org/cv5tzk
DOI: 10.1109/escience.2010.21

3. **COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project**
Frank T Bergmann, Richard Adams, Stuart Moodie, Jonathan Cooper, Mihai Glont, Martin Golebiewski, Michael Hucka, Camille Laibe, Andrew K Miller, David P Nickerson, … Nicolas Le Novère
BMC Bioinformatics (2014-12) https://doi.org/gb8wc5
DOI: 10.1186/s12859-014-0369-z · PMID: 25494900 · PMCID: PMC4272562

4. **ROHub — A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science**
Raúl Palma, Piotr Hołubowicz, Oscar Corcho, José Manuel Gómez-Pérez, Cezary Mazurek
Communications in Computer and Information Science (2014) https://doi.org/gf5m6p
DOI: 10.1007/978-3-319-12024-9_9

5. **Research Object Bundle 1.0**
Stian Soiland-Reyes, Matthew Gamble, Robert Haines
Zenodo (2014-11-05) https://doi.org/gf5m6k
DOI: 10.5281/zenodo.12586

6. **Reproducible big data science: A case study in continuous FAIRness**
Ravi Madduri, Kyle Chard, Mike D'Arcy, Segun C. Jung, Alexis Rodriguez, Dinanath Sulakhe, Eric

Deutsch, Cory Funk, Ben Heavner, Matthew Richards, … Ian Foster
PLOS ONE (2019-04-11) https://doi.org/gf5m6s
DOI: 10.1371/journal.pone.0213013 · PMID: 30973881 · PMCID: PMC6459504

7. **Datacrate Submisssion To The Workshop On Research Objects**
Peter Sefton
Zenodo (2018-07-15) https://doi.org/gf5twt
DOI: 10.5281/zenodo.1445817

8. **A lightweight approach to research object data packaging**
Eoghan Ó Carragáin, Carole Goble, Peter Sefton, Stian Soiland-Reyes
Zenodo (2019-06-20) https://doi.org/gf5twv
DOI: 10.5281/zenodo.3250687

9. **Ro-Combine-Archive**
Stian Soiland-Reyes, Matthew Gamble
Zenodo (2014-04-28) https://doi.org/gf5m6t
DOI: 10.5281/zenodo.10439

10. **Applying linked data approaches to pharmacology: Architectural decisions and implementation**
Gray Alasdair J.G., Groth Paul, Loizou Antonis, Askjaer Sune, Brenninkmeijer Christian, Burger Kees, Chichester Christine, Evelo Chris T., Goble Carole, Harland Lee, … Williams Antony J.
Semantic Web (2014) https://doi.org/gf5m6j
DOI: 10.3233/sw-2012-0088

11. **Preserving Reproducibility: Provenance and Executable Containers in DataONE Data Packages**
Bryce Mecum, Matthew B. Jones, Dave Vieglais, Craig Willis
2018 IEEE 14th International Conference on e-Science (e-Science) (2018-10) https://doi.org/gf5m6q
DOI: 10.1109/escience.2018.00019

12. **CWL Viewer: the common workflow language viewer**
Mark Robinson, Stian Soiland-Reyes, Michael R. Crusoe, Carole Goble
F1000Research (2017) https://doi.org/cbq2
DOI: 10.7490/f1000research.1114375.1

13. **A workflow PROV-corpus based on taverna and wings**
Khalid Belhajjame, Jun Zhao, Daniel Garijo, Aleix Garrido, Stian Soiland-Reyes, Pinar Alper, Oscar Corcho
Proceedings of the Joint EDBT/ICDT 2013 Workshops on - EDBT '13 (2013) https://doi.org/gf5m6r
DOI: 10.1145/2457317.2457376

14. **Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv**
Farah Zaib Khan, Stian Soiland-Reyes, Richard O. Sinnott, Andrew Lonie, Carole Goble, Michael R. Crusoe
Zenodo (2019-07-15) https://doi.org/gf5tg8
DOI: 10.5281/zenodo.1208477

15. **W2Share Case Study: Workflow Research Object (Wro)**
Lucas Carvalho, Claudia Bauzer Medeiros

Zenodo (2018-10-18) https://doi.org/gf5m6m
DOI: 10.5281/zenodo.1465897

16. **CWL run of Alignment Workflow (CWLProv 0.6.0 Research Object)**
Stian Soiland-Reyes
Mendeley (2018-12-04) https://doi.org/gf5m6h
DOI: 10.17632/6wtpgr3kbj.1

17. **The Scientific Filesystem**
Vanessa Sochat
GigaScience (2018-03-13) https://doi.org/gdwq7f
DOI: 10.1093/gigascience/giy023 · PMID: 29718213 · PMCID: PMC5952957

18. **Common Workflow Language, v1.0**
Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, … Luka Stojanovic
Figshare (2016) https://doi.org/gf6ppg
DOI: 10.6084/m9.figshare.3115156.v2