From Open Citation Data to Linked Open Data: a prototype at the ERC

Diego Chialva
ERC Executive Agency
Unit A1

The European Research Council
September 4th 2018

European Commission

Outline

European Research Council
Established by the European Commission

- Open citations in the research landscape
- (Open) Citations: embedding the graph
- A prototype at the ERC
  - data modelling (ontology)
  - corpus creation
  - workflows and tools

Research as a social fact and the value of open citations

European Research Council
Established by the European Commission

- Citation networks provide snapshots of science in its workings (collaboration, research fields, genealogy, ...)

- But research is also a social phenomenon involving stakeholders outside the "laboratory"
  - It requires funding and support policies
  - It affects society also via its outcomes

European Commission

Research as a social fact and the value of open citations

European Research Council
Established by the European Commission

- This is also well recognised within the I4OC Recall its manifesto:

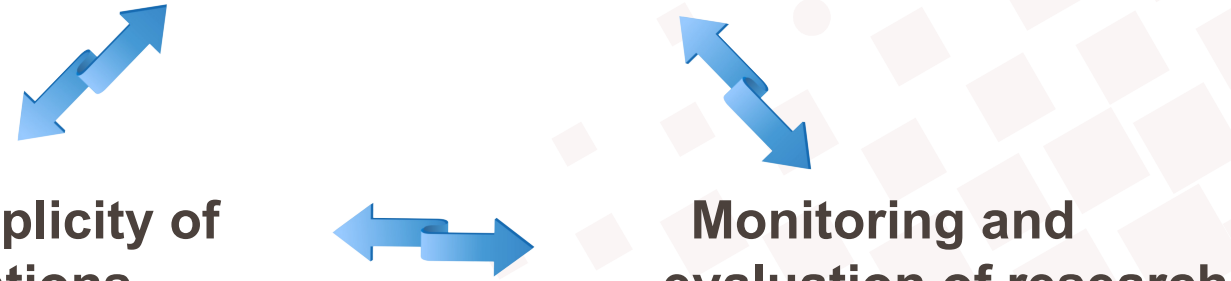
“Key benefits of achieving this aim include:

- *The ability to build new services over the open citation data, for the benefit of publishers, researchers, funding agencies, academic institutions and the general public, as well as enhancing existing services. "*

European Commission

Research as a social fact and the value of open citations

European Research Council
Established by the European Commission



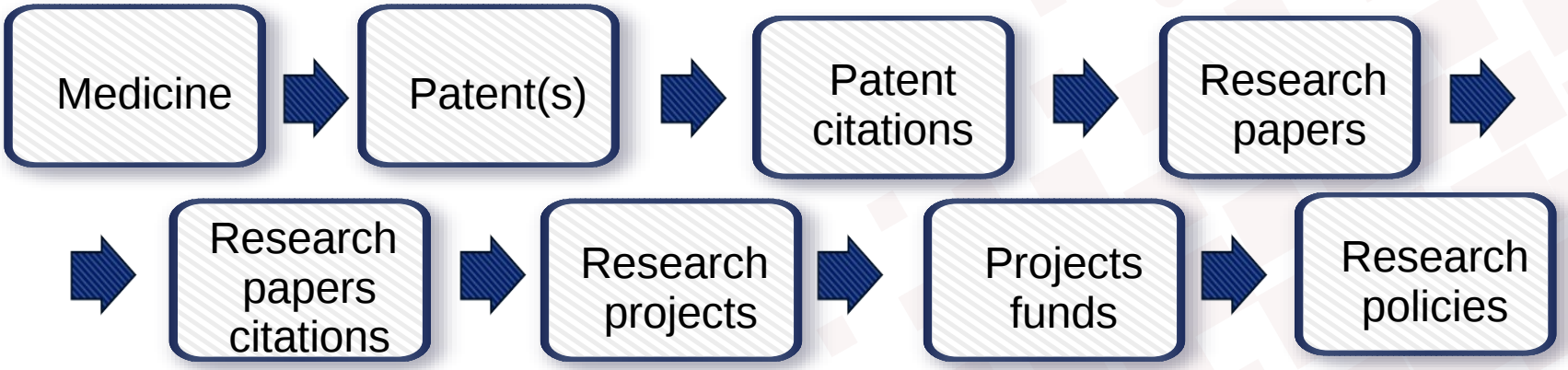
European Commission

Open citations as part of the Open Linked Data

European Research Council
Established by the European Commission

A few examples of questions:

- **what research and research policy contributed most to that medicine?**



- **what research/policy influenced most that act of legislation?**

Open citations as part of the Open Linked Data

European Research Council
Established by the European Commission

- Answering such questions (monitoring and evaluating) requires:
  - collecting and processing a large amount of data
  - working with data coming from a number of different sources
  - relationships between data forms complex networks

Open Linked Data Graph

European Commission

Open citations as part of the Open Linked Data

European Research Council
Established by the European Commission

- There are several issues:
  - processing done in isolation by each actor and on a ad-hoc basis
  - data models are not standardized → barriers to automation and re-use of data
  - data formats are not standardized → lack of interoperability.
  - lack of interoperability and data/analysis contextualisation → limits the analysis reach

European Commission

Open citations as part of the Open Linked Data

European Research Council
Established by the European Commission

- How can we create a proper Open Linked Data Base/Graph?

- Necessary steps:

1) conceptualisation and formalisation of the research landscape embedding the citation graphs → data models, vocabularies and ontologies

2) corpus creation → data recovery, data curation and reconciliation

3) establishing a workflow → toolchain definition

European Commission

1) Ontologies, vocabularies and data models

European Research Council
Established by the European Commission

- During the latest years relevant ontologies have been developed:
  - namely, the OpenCitation ontology
- But a sizable part of the landscape has remained only lightly modeled
  - the funding aspects of research and their relations with research projects

European Commission

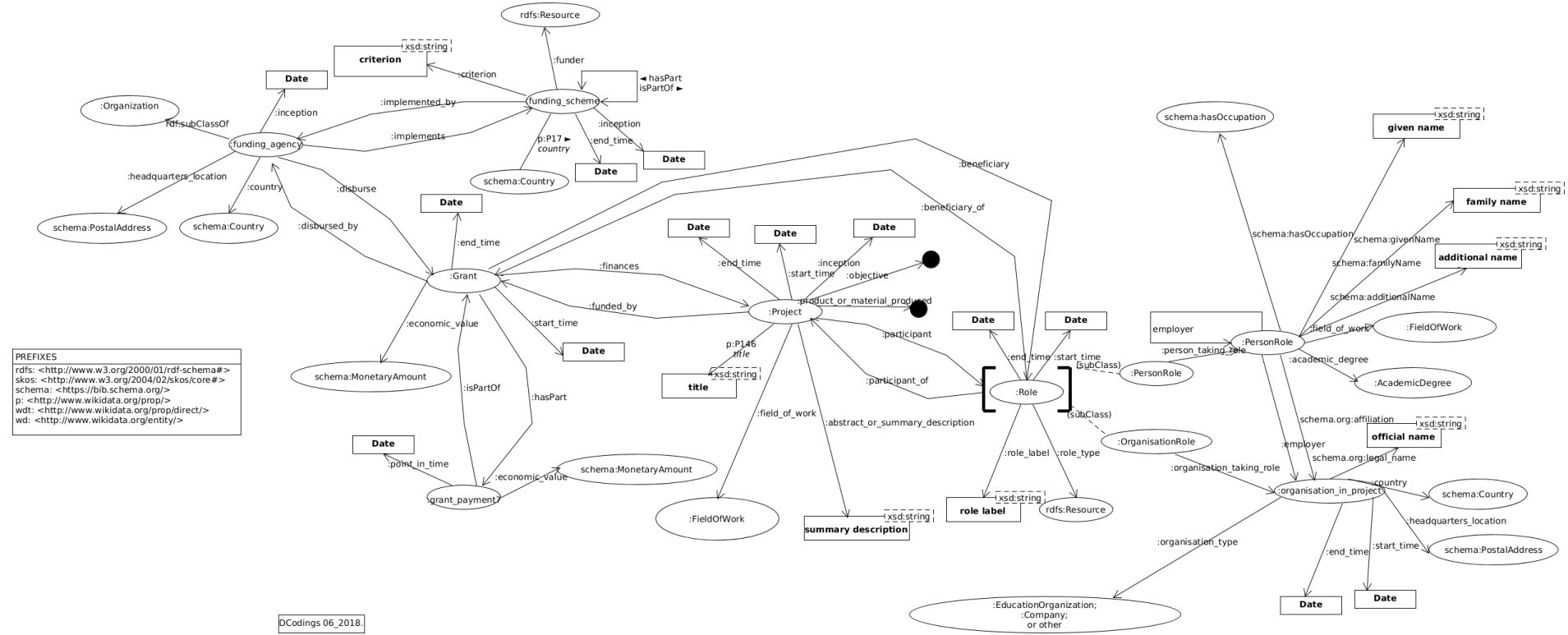
1) The DIEGO model

An initial model was presented and discussed at the Wikibase Berlin Workshop, June 2018, and dubbed by the participants the:

DIEGO = "Data Integration Extension for Grants Ontology"

- It's development has been:
  - Bottom up (data driven)
  - Consensus driven

1) The DIEGO model graph



1) The DIEGO ontology

Versions of the ontology have already been

- adapted for Wikidata (RDF-enhanced datamodel with qualifiers and references)
- proposed and presently discussed in schema.org
- Different formalisms for description and serialisations provided

1) Main principles and design decisions

- Seven principal classes: Project, Grant, FundingAgency, FundingScheme, Role, Person, Organisation
- a Project is an organised endeavour (collective or individual) planned to reach a particular aim or achieve a result
- a Grant is a disbursed fund paid to a recipient or beneficiary and the process for it
- a Project may be funded by one or more Grants simultaneously or in sequence
- a Grant may fund one or several Projects

1) Main principles and design decisions

- a Participant in a Project may not beneficiary pof the Grant(s) funding the Project
- Grants can be awarded to Person(s) or to Organisation(s)
- Projects can be participated by Person(s) or to Organisation(s)
- Funding Schemes are specifications of Grant coverage, eligibility, reimbursement rates, specific criteria for funding, population targets, and similar features
- Funding Schemes may be sub-specifications of more general Funding schemes

2) Corpus creation

European Research Council
Established by the European Commission

- **We have started with data obtainable from funding agencies**

1. Australian Research Council
2. SNSF Swiss National Science Foundation
3. Croatian Science Foundation
4. US National institute of Health
5. US National Science Foundation
6. European Union Funding - Research Framework Programme
7. Research Councils UK (RCUK)
8. grants awarded by the Europe PMC Funders

- **But the absolute lack of standards in the released data require heavy data reconciliation and preprocessing**

European Commission

2) Corpus creation

European Research Council
Established by the European Commission

- An important problem for scaling: data curation and maintenance. Possible models:
  - single repository and curator (but which authority?)
  - multiple repositories and curators (community effect? AI?)

The lack of standards and data homogeneity also causes problems to data validation

  - But positive effects could come from a shared sufficiently rich and apt model

- And for the remaining source specificities → A shared semantic graph helps: consensus model ↔ graph merging

European Commission

3) Workflows, Technologies and toolchain

- Choice of triplestore/base technology: WIKIBASE
- We developed and will make available (not as ERC tool) a "Wikibase Universal Bot" based on the WikiDataIntegrator library
  - Mitigate the Wikidata-bias and medicine-bias of WDI
  - Transparent to coding bots:
    - only data model (yaml file) necessary
    - Wikibase account and WDI write parameters specified via simple config files

Status of the project

- The modeling is complete → Publication will soon appear
- The corpus is being created (so far cleaned about 32000 project and grant records) and the Wikibase triplestore is being populated (http://orig.wmflabs.org)
- The Wikibase Universal Bot has been developed and fully tested, documentation is being completed → will soon be made publicly available (check in a while github.com/dcodings)

European Research Council
Established by the European Commission

Thank you !

European Commission