

Integration of a national e-theses online service with institutional repositories

Vasily Bunakov (STFC UKRI)
Frances Madden (British Library)



Open Repositories 2019, Hamburg, 10-13 June 2019

FREYA in a nutshell

- FREYA is a Horizon 2020 project (grant agreement no. 777523)
- FREYA is about persistent identifiers and connections between them
 - “... iteratively extend a robust environment for Persistent Identifiers (PIDs) into a core component of European and global research e-infrastructures”
- Builds on THOR (which in turn built on ODIN)

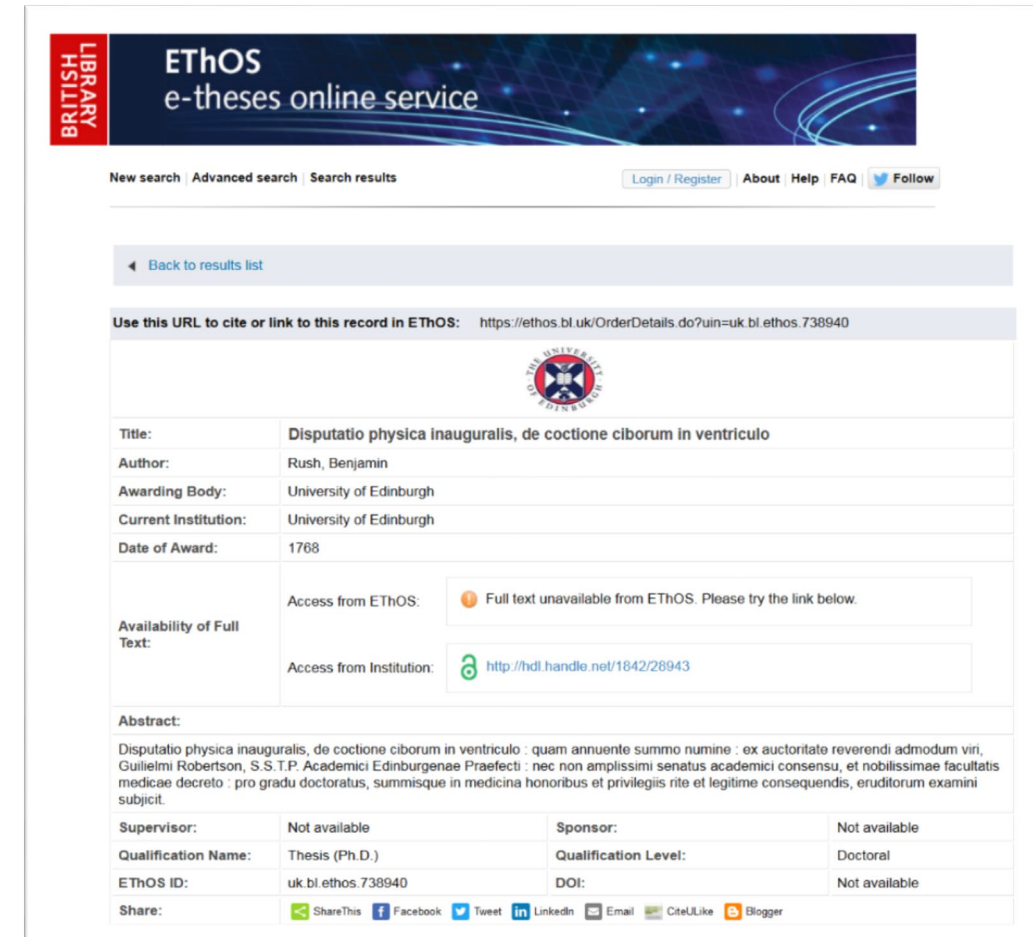
www.project-freya.eu



PID Forum: www.pidforum.org

EThOS repository at the British Library

- Index of UK theses dating back to 1768
- Contains 500k+ records
- Mixture of metadata only, full text in institutional repositories, full text held in EThoS
- Records harvested by OAI-PMH from institutional repositories
- Supports PIDs
 - ISNIs assigned to all thesis authors by the BL
 - DOIs supported where provided
 - Each record has an EThOS ID
- <https://ethos.bl.uk/>



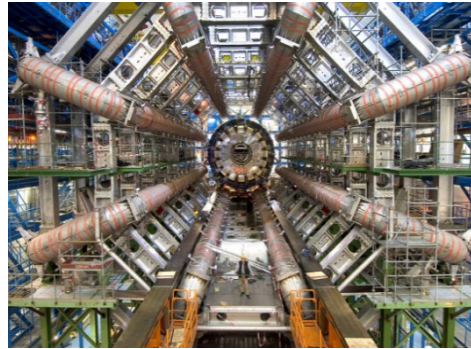
The screenshot displays the EThoS e-theses online service interface. At the top, the British Library logo is visible on the left, and the service name 'EThoS e-theses online service' is centered. Navigation links for 'New search', 'Advanced search', and 'Search results' are present, along with 'Login / Register', 'About', 'Help', 'FAQ', and 'Follow' buttons. A 'Back to results list' link is located below the search area. The main content area features a citation URL: 'Use this URL to cite or link to this record in EThoS: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.738940'. Below this is the University of Edinburgh logo. A metadata table follows, listing details such as Title, Author, Awarding Body, Current Institution, and Date of Award. The 'Availability of Full Text' section indicates that full text is unavailable from EThoS but provides a link to the institutional repository. An abstract is provided in Latin. At the bottom, a table lists Supervisor, Sponsor, Qualification Name, Qualification Level, EThoS ID, and DOI. A 'Share' section at the very bottom offers various social media and email sharing options.

Title:	Disputatio physica inauguralis, de coctione ciborum in ventriculo		
Author:	Rush, Benjamin		
Awarding Body:	University of Edinburgh		
Current Institution:	University of Edinburgh		
Date of Award:	1768		
Availability of Full Text:	Access from EThoS:	Full text unavailable from EThoS. Please try the link below.	
	Access from Institution:	http://hdl.handle.net/1842/28943	
Abstract:	Disputatio physica inauguralis, de coctione ciborum in ventriculo : quam annuente summo numine : ex auctoritate reverendi admodum viri, Guilelmi Robertson, S.S.T.P. Academici Edinburgenae Praefecti : nec non amplissimi senatus academici consensu, et nobilissimae facultatis medicae decreto : pro gradu doctoratus, summisque in medicina honoribus et privilegiis rite et legitime consequendis, eruditorum examini subjicit.		
Supervisor:	Not available	Sponsor:	Not available
Qualification Name:	Thesis (Ph.D.)	Qualification Level:	Doctoral
EThoS ID:	uk.bl.ethos.738940	DOI:	Not available
Share:	ShareThis Facebook Tweet LinkedIn Email CiteULike Blogger		

Science and Technology Facilities Council and its research facilities

STFC funds and operates large scale instruments for the UK and visitor researchers in:

- physics, astronomy
- chemistry, materials
- biology, medicine



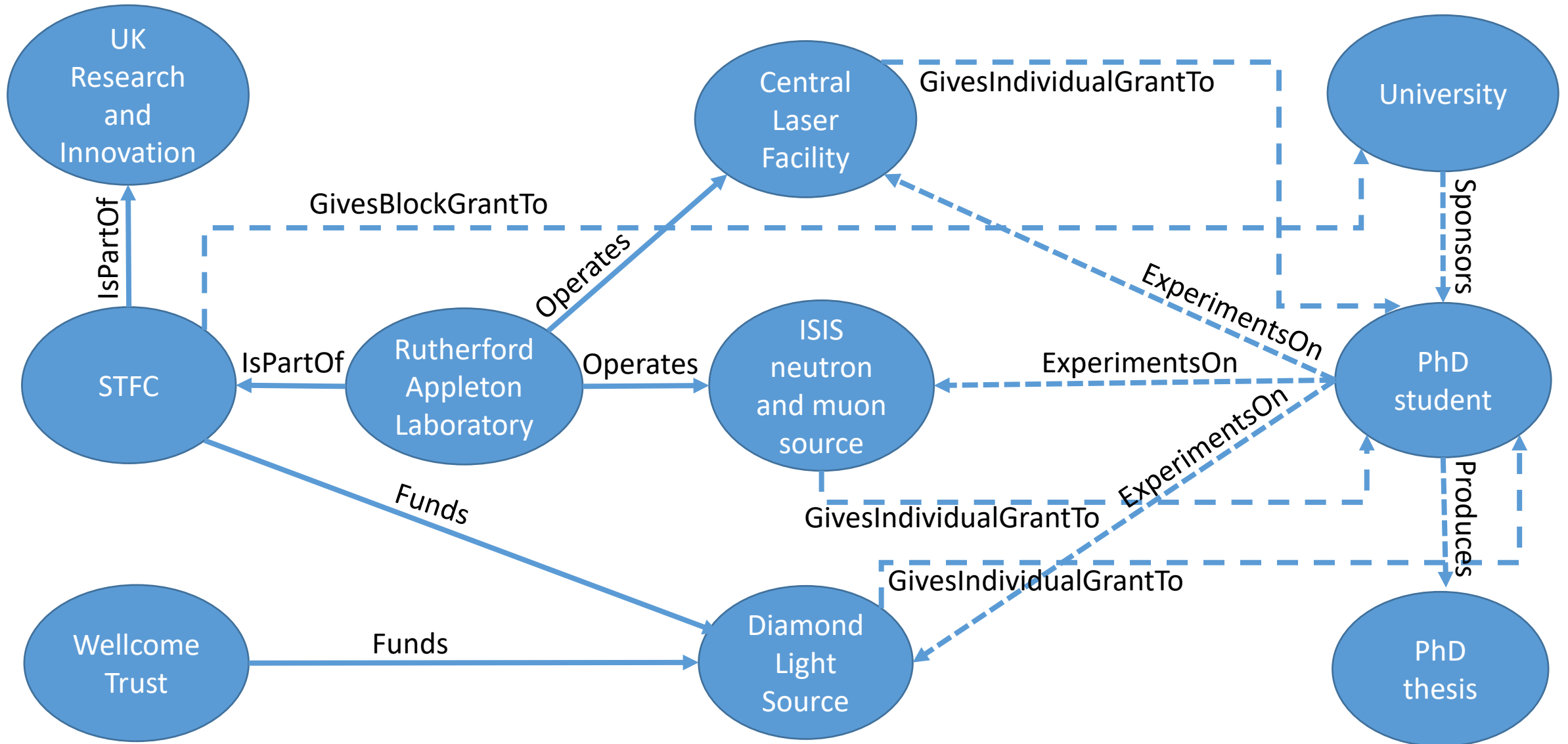
STFC research facilities:

- ISIS neutron and muon source
www.isis.stfc.ac.uk
- Central Laser Facility
www.clf.stfc.ac.uk
- Diamond Light Source
(co-owned by STFC and Wellcome Trust)
www.diamond.ac.uk

Why the PhDs use case is important for STFC

- STFC is a funder of PhDs
- ISIS, CLF and Diamond are funders-in-kind, also direct (monetary) funders in some cases
- A good case for STFC Open Science
- Good habits like giving proper attribution to facilities could be better adopted if introduced through young researchers

Organizational, operational and funding context of the PhD research supported by STFC

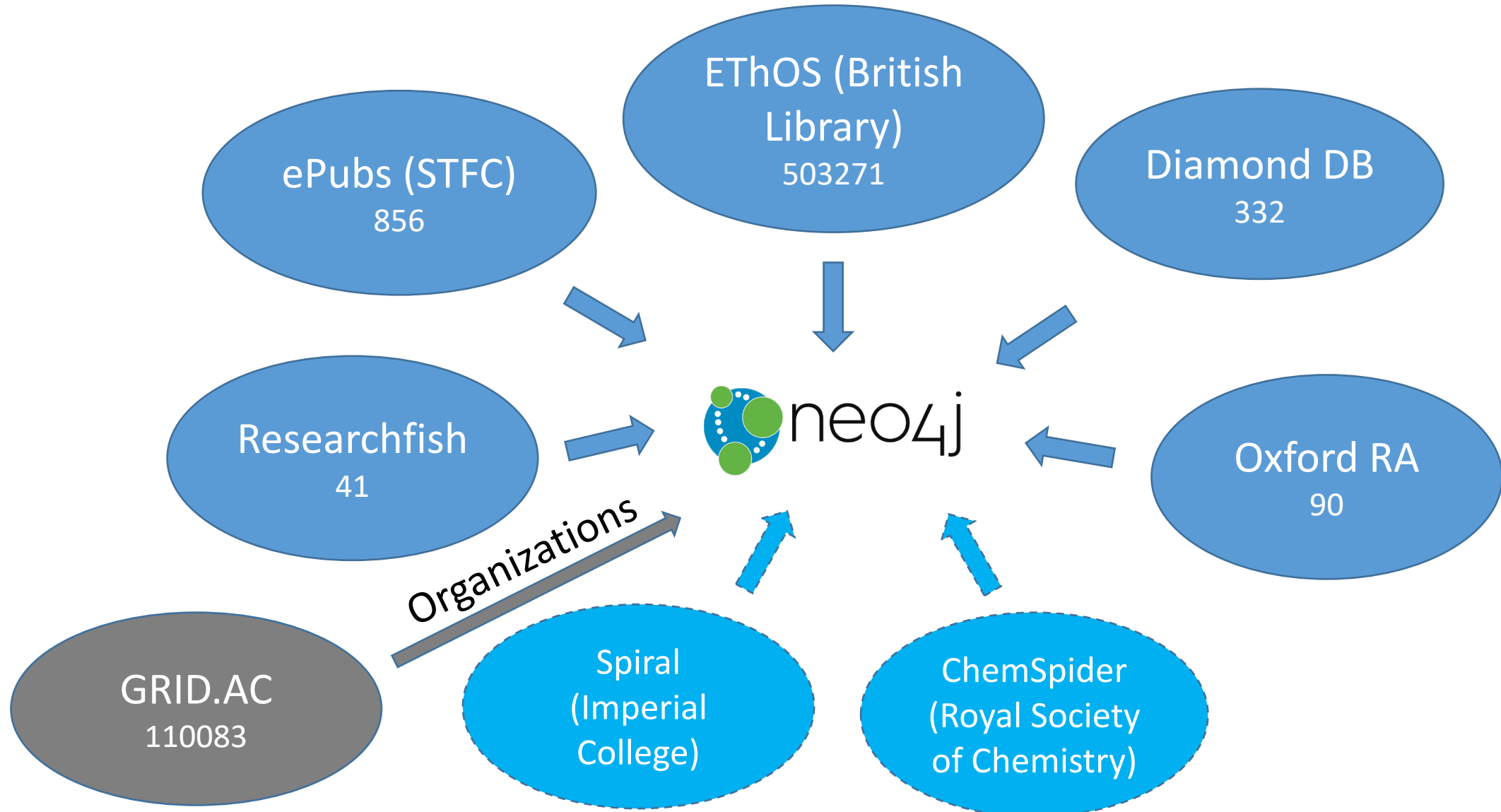


Why the PhDs use case is important for FREYA

- Collaboration: British Library and STFC are the FREYA partners and operate repositories that can be used for data integration
- Validation of new PID services – for Organizations and Instruments – and supplying feedback for their improvement
- Demonstration of PID graph value in a disciplinary context
- Integration of a disciplinary graph in a common PID graph via reasonable interfaces
- Most generic goal: contribution to and promotion of European Open Science Cloud (EOSC)

How do we build the graph?

Data sources



Why we need fuzzy matching: Examples of the same PhD theses in Oxford repository and in EThOS

ox.ID	ox.Title	ox.Authors	ox.Year	bl.Title	bl.Author	bl.Date	bl.URL
uuid:ab468708-6c14-4381-8afb-9d0f3b26ca85	Determination of the CKM phase γ at LHCb using the decay mode B_{\pm} to DK_{\pm} and a study of the decays D^0 to $KS^0K^{\pm}\pi^{\mp}$ using data from the CLEO experiment	S. Malde,G. Wilkinson,Daniel Johnson	2013	Determination of the CKM phase γ at LHCb using the decay mode B_{\pm} to DK_{\pm} and a study of the decays D^0 to $KS^0K^{\pm}\pi^{\mp}$ using data from the CLEO experiment	Johnson, D.	2013	http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.595983
uuid:181c28c2-121a-46f6-baac-c45209f7cc4a	Measurement of the inclusive $W^{+/-}$ cross section at $(\text{sq.root})s = 7$ TeV with the ATLAS detector	Adrian Lewis,Jeff Tseng	2013	Measurement of the inclusive $W^{+/-}$ cross section at $\sqrt{s} = 7$ TeV with the ATLAS detector	Lewis, Adrian	2013	http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.627800
uuid:25b20fa4-8e79-43b9-83de-225f17e333ea	Searches for new physics using Dijet Angular Distributions in proton-proton collisions at $\sqrt{s} = 7$ TeV collected with the ATLAS detector	Ryan Mark Buckingham ,Cigdem Issever	2013	Searches for new physics using Dijet Angular Distributions in proton-proton collisions at $\sqrt{s} = 7$ TeV collected with the ATLAS detector	Buckingham, Ryan Mark	2013	http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.581349

Choosing the optimal distance threshold

Scope of experiment: 58 records in ePubs versus 12049 in EThOS attributed to year 2017

Threshold for Levenshtein distance)* between ePubs and EThoS titles	Number of matches by the algorithm	True positive matches	False positive matches
5	11	11	0
10	15	15	0
15	16	16	0
20	16	16	0
25	30	16	14

15 turns out to be a reasonable threshold that allows to capture all true positives and does not result in false positives

Occasional false positives still happen at 15 characters threshold:

“Lattice dynamics in materials for energy applications” in ePubs was falsely matched with

“Lead-based materials for energy applications” in EThOS (this was 1 false versus 44 true matches for Year 2015)

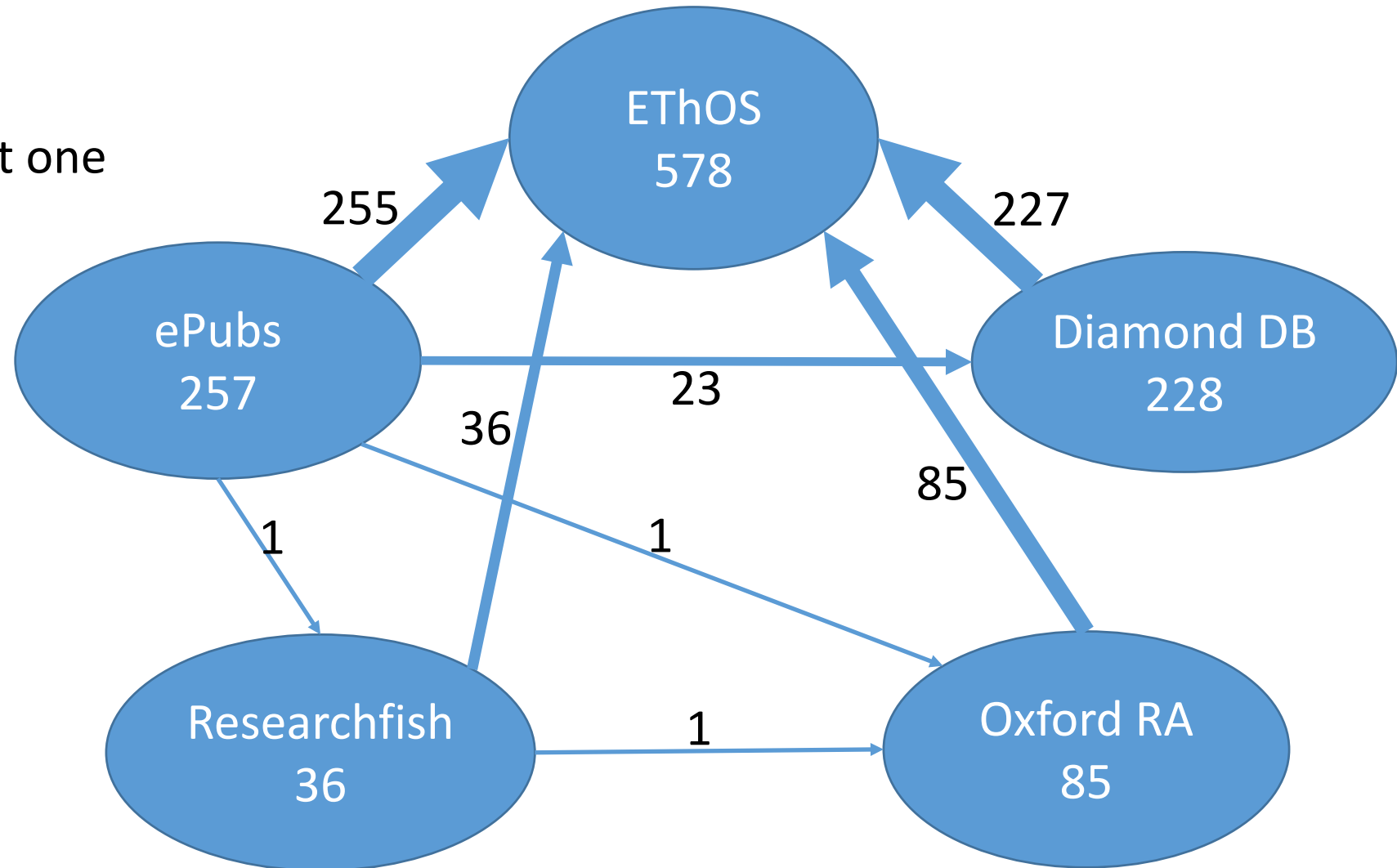
)* Minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other, see https://en.wikipedia.org/wiki/Levenshtein_distance

Only related nodes (that represent repository records) with counts of relations created

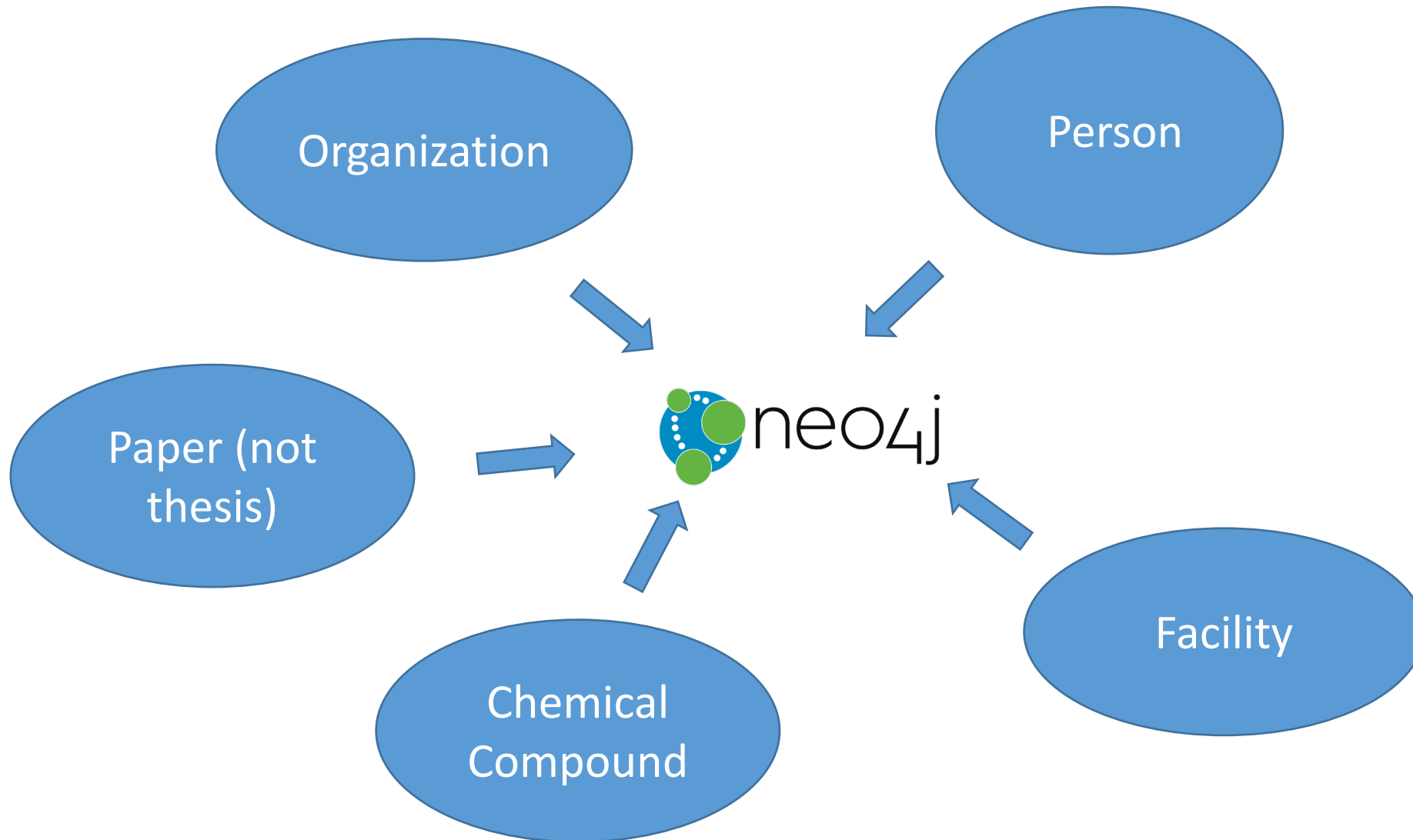
● 1184 nodes
having at least one
relation

● ● 629
paired
nodes

● ● ● 48
tripled
nodes



More node types created



Relations created

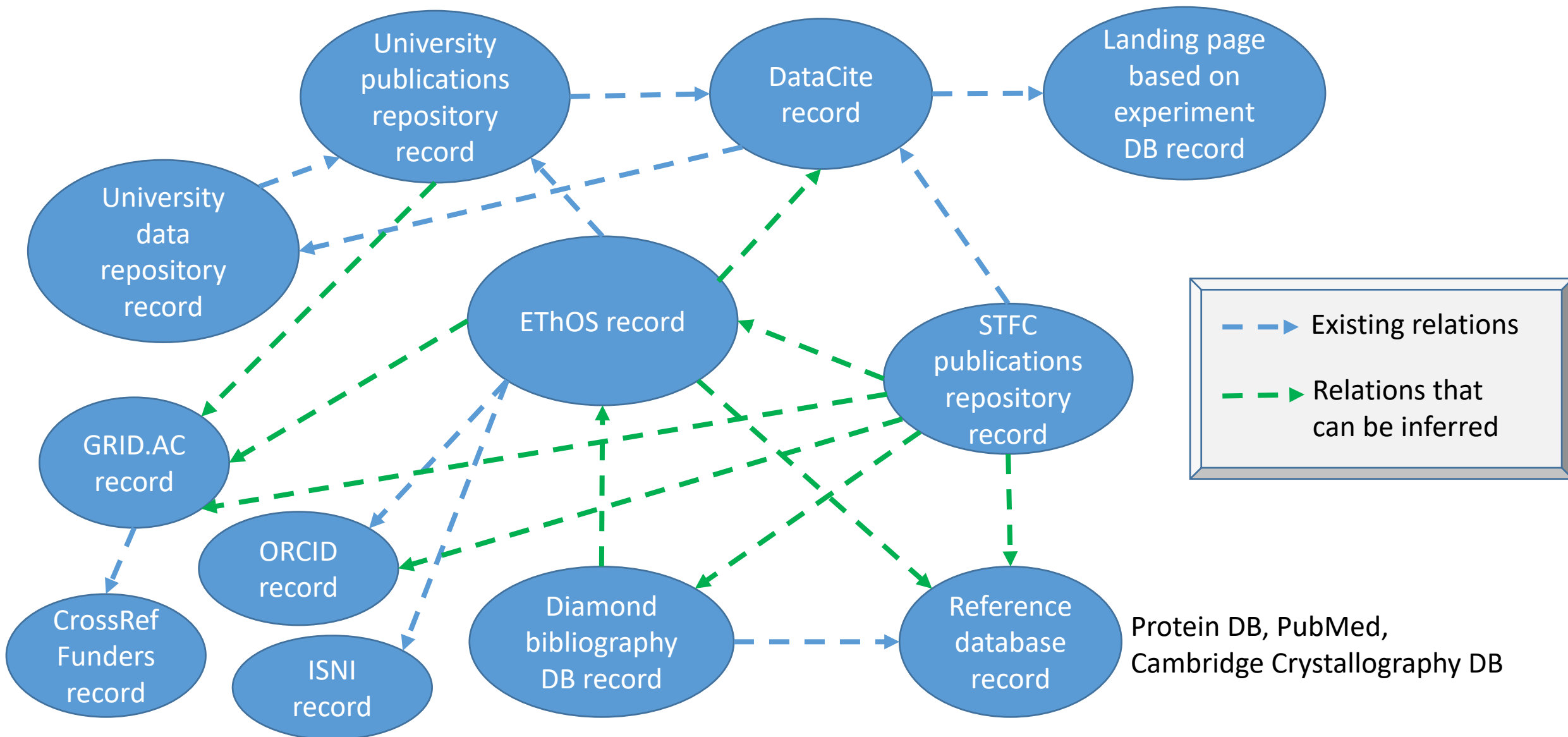
Relations created	Relations meaning	Numbers
AwardedDegreeTo	Connects University and PhD awarded with the degree	1752
Authored	Connects a PhD and a thesis that she authored	1746
sameThesisAs	Connects different manifestations of the same PhD thesis	1262
ExperimentedOn	Connects a PhD and a facility she experimented on	924
Sponsored	Connects a PhD and a funder who sponsored her	576

Imperial College PhDs who experimented on STFC facilities



How the graph can be used

Repositories perspective: what can be linked to what



Enrichment and harmonization of records as a challenge (and an incentive) for building a knowledge graph with as much use of PIDs as possible

MATCH (ethos:ETHOS_Thesis)-[r:sameThesisAs]-(x) WHERE ethos.Funders IS NULL RETURN count(ethos)	MATCH (ethos:ETHOS_Thesis)-[r:sameThesisAs]-(x) WHERE ethos.Funders IS NOT NULL RETURN count(ethos)
454	150

Cases where STFC sponsored a PhD research (via monetary funding or via facilities' grants-in-kind) but EThOS "Funders" is empty

Not all of these EThOS records connected to STFC or Diamond repository records and where "Funders" is not NULL actually mention STFC or Diamond as a Funder

And where they do mention STFC as a funder, another issue is observed: as EThOS "Funders" is currently a free-text, STFC can be referred to as:

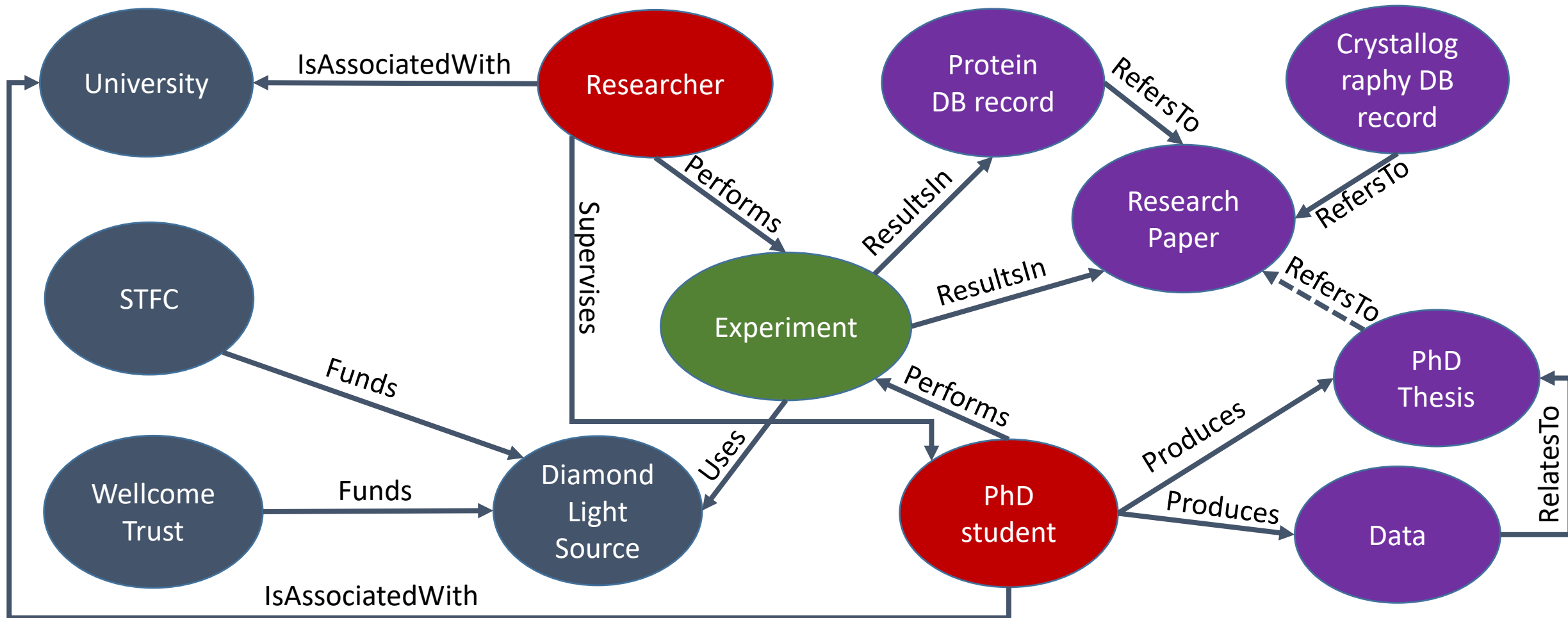
- "Science and Technology Facilities Council (STFC)"
- "Science and Technology Facilities Council"
- "Science & Technology Facilities Council"
- "Science and Technology Facilities Council (Great Britain) (STFC)"
- "STFC"

Previous slide was about what EThOS can get from the graph:

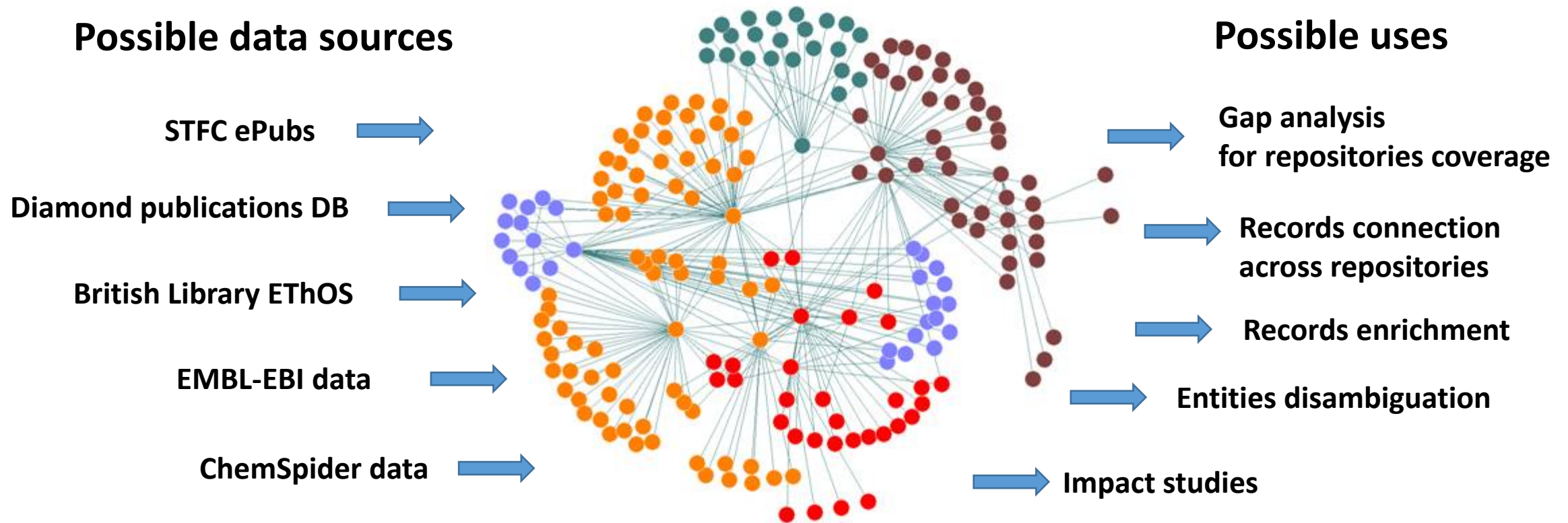
- a) more records clearly attributed to STFC as a sponsor of PhD research,
- b) STFC name uniformed across all records.

Yet STFC repositories can be enriched using the same graph, too, as it contains theses nodes attributed to STFC only by EThOS, not by any of the STFC repositories.

Enrichment and harmonization of repository records is a decent but a “traditional” goal. More ambitious and “modern” goal is building and exploiting a knowledge graph as a new multi-purpose Research Information Management infrastructure.



Support of impact studies is not the only purpose,
also PhD theses records can be just a “seed” of a larger graph.
PID graph is a (new kind of) infrastructure for Open Science



Thank you!



Web: www.project-freya.eu
Email: info@project-freya.eu
Twitter: [@freya_eu](https://twitter.com/freya_eu)

PID Forum: www.pidforum.org



The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523