

Public Knowledge Graphs

Industry Connections Activity Initiation Document (ICAID)

Version: 2.0, 14 May 2021

IC19-005-02 Approved by the IESS SMDC 14 June 2021

Instructions

- Instructions on how to fill out this form are shown in red. It is recommended to leave the instructions in the final document and simply add the requested information where indicated.
- **Shaded Text** indicates a placeholder that should be replaced with information specific to this ICAID, and the shading removed.
- Completed forms, in Word format, or any questions should be sent to the IEEE Standards Association (IEEE SA) Industry Connections Committee (ICCom) Administrator at the following address: industryconnections@ieee.org.
- The version number above, along with the date, may be used by the submitter to distinguish successive updates of this document. A separate, unique Industry Connections (IC) Activity Number will be assigned when the document is submitted to the ICCom Administrator.

1. Contact

Provide the name and contact information of the primary contact person for this IC activity. Affiliation is any entity that provides the person financial or other substantive support, for which the person may feel an obligation. If necessary, a second/alternate contact person's information may also be provided.

Name: Samuel Klein
Email Address: meta.sj@gmail.com
Employer: Knowledge Futures Group
Affiliation: Wikidata, Underlay Project

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

2. Participation and Voting Model

Specify whether this activity will be entity-based (participants are entities, which may have multiple representatives, one-entity-one-vote), or individual-based (participants represent themselves, one-person-one-vote).

Individual-Based

3. Purpose

3.1 Motivation and Goal

Briefly explain the context and motivation for starting this IC activity, and the overall purpose or goal to be accomplished.

Motivation:

- Knowledge graphs constructed from public data about the world, including the outputs of scientific research and public collaborations, and world-models and algorithms derived from them, are important public resources. They benefit from being shared, and transparently versioned and sourced. However in most cases the public versions of these graphs and algorithms are incomplete, while many competing and redundant closed versions exist.
- In particular, the reliability and provenance of information is often hidden and bundled into a single assessment of reliability of the provider. Reliability of information is more effectively assessed when the information is read and used, benefiting from hindsight and context of use. Trying to establish the reliability of assertions when they are recorded is contrary to the principles of testing and fallibility.

We want to identify and support the creation of open, federated graphs of knowledge, using available protocols for storage and mirroring, alignment of different graphs, clustering and disambiguation, annotation, tracing and adding provenance. Separating the storage of knowledge + its known provenance + its implications about the world, from its inferred provenance and implications, and the evaluations of its truth or reliability by others.

We will focus on public knowledge and public interfaces, and on machine-mediated access and knowledge representations, for constructing world and local models for machine learning.

Goals:

- Identifying existing knowledge bases and methods for automating constructions of new ones. Identifying existing protocols, and proposing new variations, to fill gaps needed to represent this knowledge. identifying protocols and tools from the decentralized web for storing and accessing such graphs.
- Identifying partners and mechanisms for documentation and maintenance.
- Identifying individual data layers and graphs that are in demand, or that exist in partial form but need completion to be widely useful.
- Defining consortia of support for both construction and maintenance of the core knowledge layers identified above.

- Providing training for researchers and data holders to use these tools to construct, publish, and share their own knowledge graphs, and to align them with large central graphs such as those maintained by Wikidata
- Organizing technical workshops to share results and research papers.

3.2 Related Work

Provide a brief comparison of this activity to existing, related efforts or standards of which you are aware (industry associations, consortia, standardization activities, etc.).

There is overlap with Open Data, Open Government, and Big Data. There is some overlap with distributed ledgers such as the Graph Protocol.

<https://thegraph.com/>

Open Data may develop related standards, and Government data sets are some of the best suited to these initial efforts.

P2807

3.3 Previously Published Material

Provide a list of any known previously published material intended for inclusion in the proposed deliverables of this activity.

There is literature about automated knowledge base construction and wikidata/dbpedia about Freebase during its development, and in the literature of technical archives since then. Participants have written more about this in the past year (1) and members of the ShEx community have joined who have an existing draft specification (2).

3.4 Potential Markets Served

Indicate the main beneficiaries of this work, and what the potential impact might be.

The primary beneficiaries are active users of current knowledge graphs at scale: research-heavy fields such as science and engineering; knowledge-heavy fields such as search and discovery; and governments and corporations that regularly mine large datasets for understanding.

3.5 How will the activity benefit the IEEE, society, or humanity?

The intended outcomes would benefit the world by providing flexible, repurposable tools and protocols, by accelerating knowledge base construction, and by making existing knowledge bases more understandable by machine-learning systems. The activity would stimulate new interest in such collaborations and may help plan workshops linking theoretical models to industry implementations.

Protocols for client-side knowledge graph compilation can allow individuals to control their information and knowledge feeds, limiting the ability of intermediaries such as publishers or search engines to manipulate their attention.

4. Estimated Timeframe

Indicate approximately how long you expect this activity to operate to achieve its proposed results (e.g., time to completion of all deliverables).

Expected Completion Date: 06/2023

IC activities are chartered for two years at a time. Activities are eligible for extension upon request and review by ICCOM and the responsible committee of the IEEE SA Board of Governors. Should an extension be required, please notify the ICCOM Administrator prior to the two-year mark.

5. Proposed Deliverables

Outline the anticipated deliverables and output from this IC activity, such as documents (e.g., white papers, reports), proposals for standards, conferences and workshops, databases, computer code, etc., and indicate the expected timeframe for each.

- At least one approved proposal for a standard describing shapes of schemas and property graphs. Associated pre-standards work to explore the idea of migrating existing work on specifications for data-shapes to IEEE to be part of such a proposal.
- Expanding our participant network to include all continents and one or more mainstream graph services providers
- A map of reference implementations of the core components of a global federated graph, including entity reconciliation, schema alignment, and namespaces for vocabularies, schemas, and transformations
- A template for identifying and maintaining a public graph using both centralized- and decentralized-web platforms.

5.1 Open Source Software Development

Indicate whether this IC Activity will develop or incorporate open source software in the deliverables. All contributions of open source software for use in Industry Connections activities shall be accompanied by an approved IEEE Contributor License Agreement (CLA) appropriate for the open source license under which the Work Product will be made available. CLAs, once accepted, are irrevocable. Industry Connections Activities shall comply with the IEEE SA open source policies and procedures and use the IEEE SA open source platform for development of open source software. Information on IEEE SA Open can be found at <https://saopen.ieee.org/>.

Will the activity develop or incorporate open source software (either normatively or informatively) in the deliverables?: **Yes**

6. Funding Requirements

Outline any contracted services or other expenses that are currently anticipated, beyond the basic support services provided to all IC activities. Indicate how those funds are expected to be obtained (e.g., through participant fees, sponsorships, government or other grants, etc.). Activities needing substantial funding may require additional reviews and approvals beyond ICom.

Participants are building shared data hosting and related tools, in addition to standards for openness, for which they are finding their own funding. The development of standards proposals is not expected to incur significant expenses.

7. Management and Procedures

7.1 Activity Oversight Committee

Indicate whether an IEEE Standards Committee or Standards Development Working Group has agreed to oversee this activity and its procedures.

Has an IEEE Standards Committee or Standards Development Working Group agreed to oversee this activity? No

If yes, indicate the IEEE committee's name and its chair's contact information.

IEEE Committee Name: None (ICom)

Chair's Name: Full Name

Chair's Email Address: who@where

Additional IEEE committee information, if any. Please indicate if you are including a letter of support from the IEEE Committee that will oversee this activity.

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

7.2 Activity Management

If no Activity Oversight Committee has been identified in 7.1 above, indicate how this activity will manage itself on a day-to-day basis (e.g., executive committee, officers, etc).

The Activity will be managed by an executive committee as defined in the Activity's Policies and Procedures.

7.3 Procedures

Indicate what documented procedures will be used to guide the operations of this activity; either (a) modified baseline *Industry Connections Activity Policies and*

Procedures, (b) Standards Committee policies and procedures accepted by the IEEE SA Standards

Board, or (c) Working Group policies and procedures accepted by the Working Group's Standards Committee. If option (a) is chosen, then ICCOM review and approval of the P&P is required. If option (b) or (c) is chosen, then ICCOM approval of the use of the P&P is required.

The Activity will follow a modified version of the Industry Connections Activity Baseline Policies and Procedures

8. Participants

8.1 Stakeholder Communities

Indicate the stakeholder communities (the types of companies or other entities, or the different groups of individuals) that are expected to be interested in this IC activity, and will be invited to participate.

Patent researchers studying inventions + products
 Economists studying the development of ideas and innovation
 Free knowledge + Open Access researchers using the scholarly + patent citation graph
 University department chairs and provosts relying on such tools for self-evaluation
 Librarians who acquire access to such data for their patrons
 Developers of personal knowledge management tools, with local graphs of context

8.2 Expected Number of Participants

Indicate the approximate number of entities (if entity-based) or individuals (if individual-based) expected to be actively involved in this activity.

About 25 in the coming year

8.3 Initial Participants

Provide a number of the entities or individuals that will be participating from the outset. It is recommended there be at least three initial participants for an entity-based activity, or five initial participants (each with a different affiliation) for an individual-based activity.

Use the following table for an entity-based activity:

Entity	Primary Contact	Additional Representatives
Entity Name	Contact Name	Name

Use the following table for an individual-based activity:

Individual	Employer	Affiliation
James Ross	U.Chicago	Booth School of Business
Samuel Klein	KFG	Wikipedia
Osmat Jefferson	QUT	Lens.org
James Weis	MIT Media Lab	KFG
Andra Waagmeester		Wikidata, ShEx dev team
Eric Prud'hommeaux	MIT CSAIL	ShEx dev team

8.4 Activity Supporter/Partner

Indicate whether an IEEE committee (including IEEE Societies and Technical Councils) has agreed to participate or support this activity. Support may include, but is not limited to, financial support, marketing support and other ways to help the Activity complete its deliverables.

Has an IEEE Committee agreed to support this activity? No

If yes, indicate the IEEE committee's name and its chair's contact information.

IEEE Committee Name: Committee Name

Chair's Name: Full Name

Chair's Email Address: who@where

Please indicate if you are including a letter of support from the IEEE Committee.