One of our clients is a major children's hospital in the Midwestern United States. Apart from being a large medical service provider, the institution is also a major research center with several hundred faculty members engaged in a variety of scientific activities beyond their medical duties.

Between them, the faculty members have published more than ten thousand peer-reviewed scientific articles, many in top ranking Pediatrics journals. Many of them regularly speak at research conferences; participate in major research projects and in various stages of major clinical trials; receive awards and citations for their work; and collaborate with other research institutions across the globe.

Tracking these activities is key to assessing the impact of the institution on the current state of medical research but combining such a variety of data sources is also a surprisingly challenging task.

The Stumbling Blocks of Tracking Research Activities

Tracking research impact, even on a relatively small scale is far from trivial. The initial issue for the children's hospital was the fragmented nature of the available information. Each person is an employee, a doctor, an author of peer-reviewed publications, an investigator in clinical trials and a public figure participating in conferences and receiving recognition for their scientific work.

The hospital (and many other Healthcare institutions like it) keeps the data in various systems where each serves the specific needs of a different department and there is no unified access or identification of individuals between databases.

In fact, the challenge is even more complicated than it appears at a first glance because the methods for data collection lead to differences in coverage and reliability. While data for activities in which the institution is directly involved (such as clinical trials, for example) is kept internally, other key information (such as records about the different researcher's publications) is gathered using self-reporting protocols. The institution runs campaigns where individuals are encouraged to fill in a form with their different publications during the reporting period.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

This approach is sub-optimal for a variety of reasons:

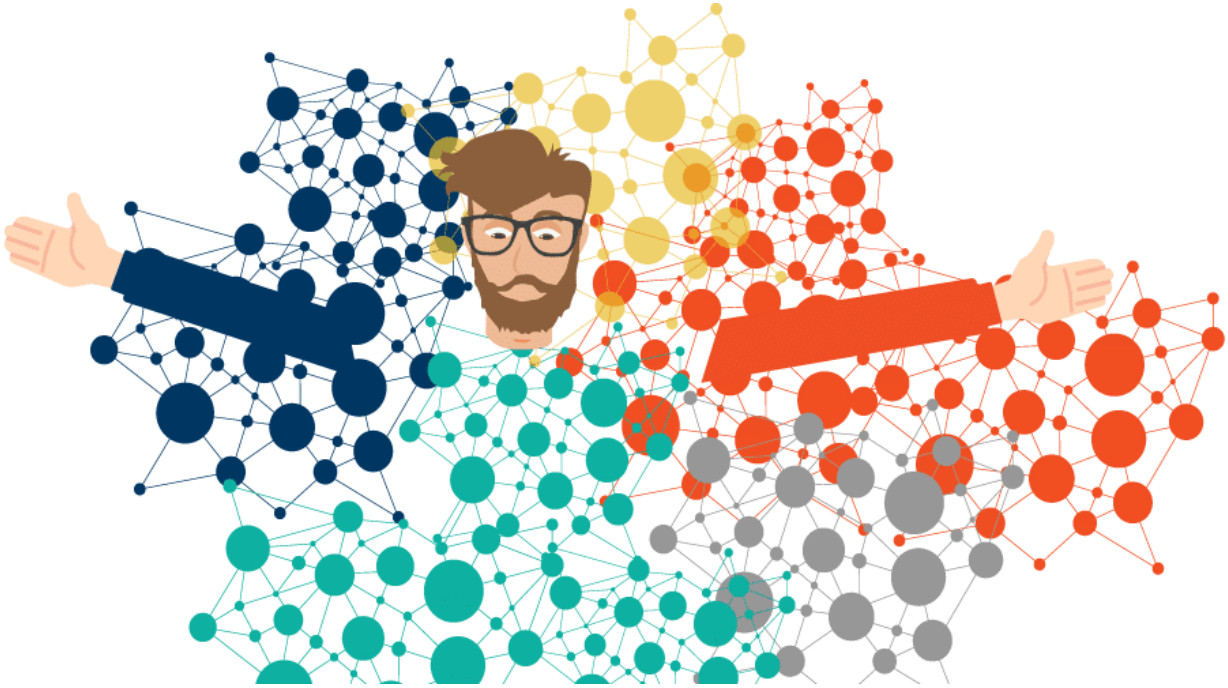
- The data is incomplete because there is no clear protocol on what should be reported.
- Not everyone complies with the existing protocol and completes the surveys, which leads to fragmented data.
- Data collected in this way can be very noisy as, for example, some researchers only report the title of their publication.
- It's surprisingly difficult to collate data between such campaigns and keep a consistent chronological record of publications.
- The process generates extra workload for faculty members who are also among the most valuable employees.

On top of the difficulties related to the protocol, there are some inherent complexities of data about academic publications. To be able to extract insights from such data, one needs access to a surprising amount of structured information:

- Even simple impact metrics such as the H-index need information at at least two degrees of separation from the researcher [1]. More modern metrics such as the academic rank calculated by Microsoft Academic rely ultimately on the structure of the entire citation network to produce a number.
- Custom metrics are no different. For example, the hospital is interested in federal and international collaborations. To be able to extract such information for a given person, one needs not only to identify all co-authors (a difficult task in itself) but to know each of their affiliations and the location of the corresponding institutions.
- Even tracking seemingly direct features such as identifying publications that received recognition in the form of citations or discussion on prestigious blogs, requires integrating several datasets, both internal and external in order to make the proper connections.

These kinds of analytics are certainly possible and can provide valuable insight into the activities of a faculty. But carrying them out by hand requires a lot of effort by people closely familiar with the field and there is still a significant risk of missing some interesting connections. A knowledge graph that integrates the several internal and some large external data sources would be able to automatically find many of these connections and free up these domain experts to explore and analyze the most interesting instances.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



Our Approach: Semantic Data Integration of Proprietary and Public LOD Sources into a Single Knowledge Graph

Fortunately, structured information about academic output is available more and more, both in commercial datasets and as Linked Open Data. Besides established players such as PubMed, Scopus and Web of Science, novel data providers such as Microsoft Academic Graph [2] give us access to almost all the structure we need for extracting insights from publications. The advantage of this approach is that it is quite flexible and can easily be adapted to work with any one or even multiple of these datasets depending on the needs and preferences of a client.

Besides information about publications, other relevant sources are government data providers such as ClinicalTrials.gov and the NIH register about grants in Healthcare. While none of these sources is individually all that interesting for a client's analysis, their true value shines through when they are combined and integrated with the internal data. In this way, all the relevant information they contain can be seamlessly queried and tied to internal records.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

So this is what we did in a nutshell:

1. First, we discussed the kind of analytics the children's hospital was interested in and examined the data they had available.
2. Based on that, we built a unified model for a knowledge graph that could support the analytics requirements of the hospital.
3. We took several internal datasets in different formats with drastically different information, which nevertheless concerned the same several hundred people and converted each individual dataset to RDF following the model.
4. Then, we connected the people between records in the different datasets to get a Doctor-360-degree view.
5. We pulled large quantities of focused data from several external datasets to fill out the parts of the knowledge graph that were required for the analytics but were not available in internal data.
6. Finally, we worked with the hospital to develop and fine-tune the analytics over the knowledge graph.

The Benefits Knowledge Graphs Bring to the Table

To better understand the benefits of working with this Research Impact Tracking knowledge graph of integrated data, let's consider an example. We already mentioned that the geographical distribution of faculty collaboration with external institutions was one of the important indicators of interest for the hospital.

After some fine-tuning, we decided on three levels of tracking the location of the affiliation of a given faculty member's co-authors: co-authors from within the state, co-authors from another US state and co-authors from a different country.

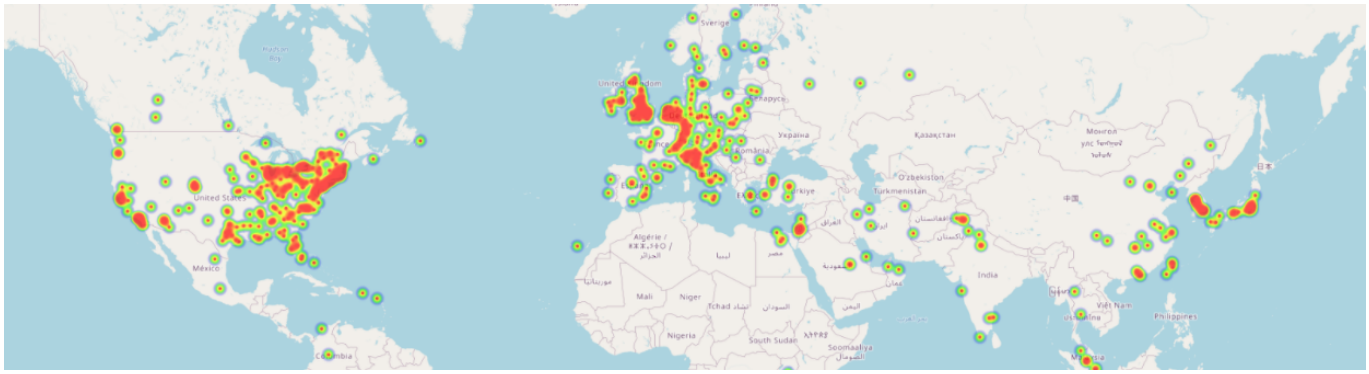
In order to compute this indicator, we need a surprising variety of data, which is not available in any single place:

- Data about the identity of the faculty members we are tracking comes from the internal faculty data of the hospital.
- Data about their publications comes not only from internal reporting but also from their publication history in PubMed and Microsoft Academic Graph.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

- Data about co-authors and their affiliations comes primarily from Microsoft Academic Graph but some missing pieces are filled in from PubMed.
- Geographical location of a co-author's affiliation – in many cases, academic datasets don't contain the exact location of a given institute so we need yet another data source – Wikidata. It enables us to collect not only the coordinates of a given institution but also its place in the administrative territorial hierarchy of its country, thus making it perfect for supporting the custom aggregation requirements of the hospital.

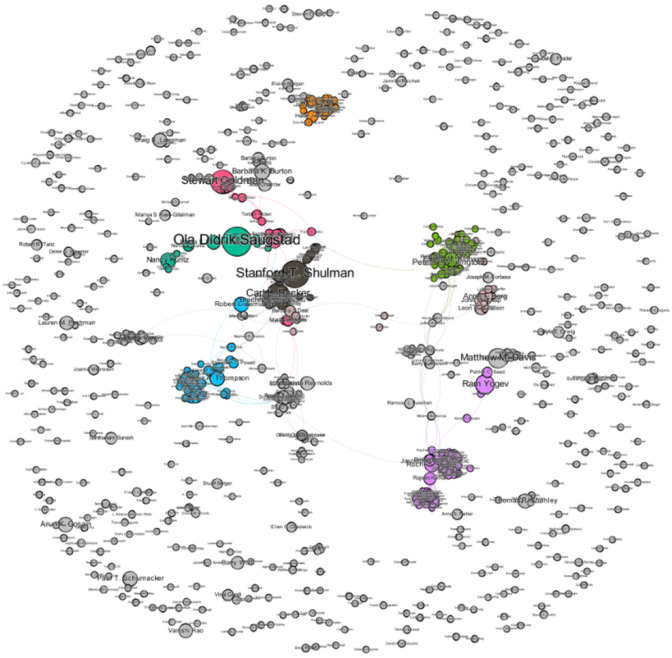
The final result looks something like this:

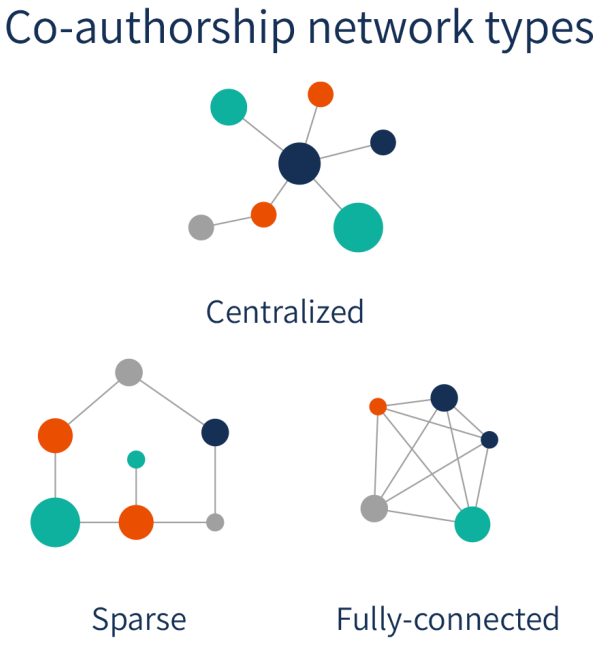


Once the data is fully integrated, it is simple to generate not only numeric reports but also much more intuitively understandable interactive visualizations.

This example is only one of many, where we need to compute custom indicators over fragmented data. We defined similar steps for analysis and calculation for other indicators of interest such as emerging co-authorship networks, publications resulting from collaboration in clinical trials, comparative rate of increase of scientific output, etc.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com





In the above image we can see a visualization that identifies the major co-authorship networks within a slice of the dataset. Through this visualization, we can see not only the most active co-authorship networks but also gain some interesting insight into their differing topologies. Some have a single central member participating in all publications, others are more fully connected, and yet others are more “stretched out” with each author collaborating with only a subset of the overall group.

Major Takeaways

As the above example shows, one of the most valuable advantages to having a rich fully integrated dataset is that many new questions and ideas come out through the analysis. Often, they can either be answered immediately or through the straightforward addition of some specific missing piece of data to the knowledge graph.

Another immediate advantage of impact metrics analytics powered by a knowledge graph is that it can enable institutions to identify trending and emerging areas of research as well as to provide easier access to public or private research grants. It can also help them distribute the financial support more wisely among top priority scientific areas and specific research

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

topics. Last but not least, it can strengthen scientific collaboration with individual researchers and institutions. As a result, large Healthcare institutions will have a much more comprehensive and up-to-date picture of their faculty research activities and will be able to make more valuable contributions to medical research.

---

[1] H-index is based on all the publications and counting their citations.

[2] Which will unfortunately be discontinued in 2022 but hopefully succeeded by https://openalex.org/

Excited by the potential of a knowledge graph based faculty research tracking?

Book a Discovery Call

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com