Not every scientists has the comfort of a well-equipped lab. However, newly available open platforms for biomedical in silico discovery could soon spark the brains of millions of researchers forming a geographically-distributed work force across the globe. This no longer requires working in a high-tech lab to contribute to the discovery of new mechanisms in health and diseases. Meanwhile, new opportunities for trainees, scientists and patients to practice annotation of genetic databases, could push the boundaries of open science towards countries where it has not yet been possible to work on such projects. In the second part of a two-part series, Barend Mons from the Leiden University Medical Centre, The Netherlands, explains how it could work in practice, and how close we are to realising this initiative.

Published in EuroScientist via SciencePOD

Euro Scientist

Biological mechanisms discovery by globally distributed research force

Case study: a new open science model applied to disease pathways discovery

A new open science approach could soon change the way we think about research. It is based on charging those keeping discoveries private to subsidise those who publicly share their findings. Here, we examine how this new way of handling scientific findings could work in practice. To illustrate this approach, we take the example of discovery of biological mechanisms underlying health and disease.

Our research with the Biosemantics Group Leiden performed in close collaboration with the Dutch bioinformatics start-up EURETOS, based in Delft, has led to the development of the privately owned EURETOS Knowledge Platform (EKP). This is a very valuable collection, or graph, of 60 million unique so-called Cardinal Assertions, derived from over 80 data sources as of January 2016. The graph features relations between two concepts—dubbed triples (including 'concept A', 'relation' and ‘concept B’)— relevant for health and diseases, such as genes, proteins, chemicals, diseases, tissues and physiological process.

By removing all redundancy from the triples, the data engineers of EURETOS have narrowed them down to Cardinal Assertions. What makes them valuable is that they are presented with their complete provenance—that is, the reference to the evidence from the scientific literature for each assertion. In turn, the linked provenance makes it possible to for humans to judge the validity of these Cardinal Assertions.

Privately versus publicly held findings

Using the EKP in a safe, fire-walled environment obviously costs money. This proprietary resource is available can be accessed subject to paying a license fee. Private clients, such as pharmaceutical companies and clinicians, would typically use it. They can safely query the database and avail of the convenience of the in-built workflows system to explore and validate or invalidate the Cardinal Assertions.

How can such resource be relevant to further open science?

A copy of the entire EKP of Cardinal Assertions is also available in an open access version for academics, provisionally called DKP. This version is accessible via a registration to enable tracking users via their ORCID number as a means to reward their contributions to open science. In fact, DKP has been established as a public-private partnership between EURETOS, and the data analysis service provided by the Dutch Techcentre for Life Sciences (DTL), based in Utrecht. By being part of this partnership, EURETOS thus supports the open science credo by providing the DKP, available free of licensee fee to DTL users.

Of course there should be mutual benefit. The academic community is expected to adopt and support the EURETOS approach, by publishing every query, annotation and addition in open access. In turn, the company, like everyone else, can use these public findings to improve their content, workflows and general service offering. Obviously, this is subject to annotators and contributors making their contribution available in open access. If they decide they want any kind of restrictions, they will have to pay and revert back to using the commercial EKP version.

New open science scenario

This is where the opportunity for developing and developed country scientists alike could arise.

Let’s take the example of Jeanine, a clinical geneticists at a large academic centre. She has a patient with Intellectual Disability (ID) and performs genome sequencing. All genes known to be associated with ID appear intact. But she notices in the sequenced genome of one of her patients that a gene called DRAXIN shows a major deletion, which may mean that it is affected in its expression and makes the wrong protein or none.

The trouble is that DRAXIN and ID do not have any direct co-occurrence in the current scientific literature. Nor in the 80+ databases encompassed in the EKP. Using classical literature search, there is no connection or obvious indirect relations. However, when looking at possible indirect associations between the gene and the disease, the EKP predicts a reasonable chance that DRAXIN could be playing a role in brain development and mental retardation.Indeed, DRAXIN appears to be strongly associated with concepts, such as forebrain development, axon chemotaxis and neutron migration. These are all concepts that can be easily connected to mental retardation and intellectual disability.

The prediction power stems from the 60 million connections stored in the database, as opposed to 8 millions connections documented by the National Center for Biotechnology Information (NCBI) database, which is not from openly available life science literature databases, such as PubMed.

Further, in the NCBI database, the abstract of the corresponding research paper does not contain the literature supporting this connection while the full text is behind a paywall. This is not a problem for jeanine because she is lucky enough to be at and institute with an expensive subscription that enables here to retrieve the full text. Herein lies the rub for open science. The very fact that Jeanine--whose activity is tracked via her ORCID number and connected to her institution--searched the EKP for the new combination of DRAXIN and Intellectual disability is a piece of information of great value.

An identical query related to a potential relation between DRAXIN and ID, could potentially be done by many independent clinical geneticists around the world.

On the record

This repeated number of searches is recorded by the system and could thus produce a signal to be investigated further by the wider scientific community. And the signal could easily picked up first by a scientist from a developing country in DKP. The nanopublication of Jeanines's first query is a way of documenting her contribution to the association via her own ORCID number and her institution. Meanwhile, there is also a time stamp associated with the query, attesting that she could be the first to think about such association

If Jeanine decides that she is quite interested in these relations, she can then automatically request, from her EKP account, a human readable article explaining the rationalisation of all indirect connections between DRAXIN and ID. Typically, these connections arise via other genes, proteins metabolites, drugs, and physiological processes, such as axon guidance and forebrain development. This advanced workflow, for instance, suggests that DRAXIN is involved in neurogenesis. However, the computer generated article does only give a very low chance (13%) to the prediction that DRAXIN may be directly involved in intellectual disability. These many indirect connections include 8 associated sub-diseases, such as Down syndrome, Backwith-Wiedemann syndrome and X-linked mental retardation.

Based on this relatively vague connection, she decides to publish the article via her chosen open access channel for anyone to read, comment on or claim. This also means that all other clinical geneticists or biologists working on DRAXIN for instance posting the same query will now see the published article and may become reviewers.

As long as this all remains in full open access, it contributes to open science and should ideally be free –namely subsidised by restricted science. In this scenario, Jeanine's goal is not to increase her impact factor but to help patients. Therefore, she does not have any desire to keep this article to herself, as sharing may help quicker discovery of more genetic causes for ID.

Like Jeanine, any researcher from across the globe, who gets really interested in this link and its rationale outlined in the article can claim the EKP/DKP computer generated association as their research hypothesis. They could then decide to do more research on it.

Should they have access to a wet laboratory, they could for example perform experiments where the DRAXIN gene is prevented from being expressed in zebrafish embryos and find that they do not develop a brain. This mean that they have contributed additional evidence to support the hypothetic link between DRAXIN and ID.

Then, they can start writing a standard article about this hypothetical relationship, with the option to include the results of their labora-tory experiments. Even if scientists do not hace ready access to wet lab facilities of expensive equipment, they could be the first claimant of the hypothesis and seek collaboration to prove it.

Should Jeanine opt retrospectively to become one of the authors of a formal publication, she is expected to contribute to the publication fee--akin to an Open science Tax--or an open access article publication fee.

Credit for contribution

Another way that developing country scientists could contribute is in annotating the database. Any triple from the EKP/DKP platform is available to be annotated by any researcher identified through their ORCID number. Annotations include both critics on existing assertions and negative results. We will soon be ready to invite millions of students and experts from around the world to annotate and improve triples with proper predicates, nuances and comments. Again if they decide to keep their annotations to themselves, they would have to pay and/or have to use the fee-based EKP.

Any open annotation using open tools like the Open RDF Knowledge Annotator (ORKA) would also be credited as one of the actions contributing to further opening science. If two concepts like a drug and a disease are co-occurring in a text and are new to EKP/DKP, according to specialised text mining tools such as Utopia Docs, they can be served up as an RDF triple in ORKA.

Students can make a name for themselves as annotators. Once more, internal annotations--keeping knowledge about certain triples private--should be paid for. We could imagine business models where annotators contributing a given number of open annotations, build up credits that allow them to do a given number of private annotations; again a model that does not put colleagues in developing countries at a disadvantage.

New business model

For this solution to work, all kinds of paid services could subsidise the free services associated with publication of findings made available for sharing and further testing. For example, users have the option to do micro-payments to preserve the confidentiality of queries in DKP. They can also choose to buy private subscriptions to EKP workflows from EURETOS. There might be many other ways of generating additional revenues.

In essence, all actions that contribute to open science in this system should be free or rewarded accordingly, potentially financially in some cases. For example, any action can be automatically recorded on its authors' CV--particularly via systems like VIVO, which recognise the unique research identification number ORCID. Thus, sharing is subsidised by not sharing; in many cases for good reasons.

I invite all scientists concerned to get together and make this happen sooner rather than later.

Barend Mons

Barend is professor in bio-semantics at the department of human genetics at the Leiden University Medical Centre (LUMC)

Disclaimer: This is my personal opinion as a scientist and cannot be quoted as the formal opinion of the High Level Expert Group for the European Open science Cloud, of which I am a member, nor attributed to LUMC, Dutch Techcentre for Life Sciences (DTL) or EURETOS. I am an independent scientific advisor to EURETOS without any financial arrangements or other forms of participation in this initiative.

Photo credit: Sabrina Campagna (CC BY-NC-ND 2.0)