Research Ideas and Outcomes 2: e11014
doi: 10.3897/rio.2.e11014

Small Grant Proposal

Shuttleworth Fellowship Application 2016

Donat Agosti ‡

‡ www.plazi.org, Bern, Switzerland

Corresponding author: Donat Agosti (agosti@amnh.org)

Reviewable v1

Received: 01 Nov 2016 | Published: 03 Nov 2016

Citation: Agosti D (2016) Shuttleworth Fellowship Application 2016. Research Ideas and Outcomes 2: e11014. https://doi.org/10.3897/rio.2.e11014

Abstract

This is the application by Donat Agosti for a Shuttleworth Fellowship with the goal of making scientific data publication an integral part of an open knowledge management system. A successful application will help move TreatmentBank from its current prototype state into a production system that converts and extracts data from taxonomic publications and makes them available as Linked Open Data. The project is about developing the infrastructure and creating a corpus of data, while at the same time serving as a showcase of open access. This is an exemplar approach that hopefully will be adopted by other scientific domains.

Keywords

fellowship, applications, Shuttleworth, biodiversity, open access, linked open data, TreatmentBank, taxonomic treatment

Personal details

Name: Agosti, Donat

Email: agosti@plazi.org

Date of Birth: 06.10.1958

Nationality: Swiss

1. Tell us about the world as you see it

The loss of biodiversity is increasingly taking a backseat to politics mainly because we still have no sound global assessment of our planet's biodiversity. To the extent this knowledge is available, it is not widely accessible. In fact, there is a stunning, almost complete disconnect between public knowledge and the scientific understanding of biodiversity published in well over 500 Million pages of highly standardized taxonomic treatments and illustrations describing the world's known 1.9M species, published by scientists from all around the world over the last 250 years. Neither do the Wikipedia and Google provide links to science-based facts underpinning global knowledge of biodiversity nor do the scientists incorporate such links. In fact, they couldn't even if they wanted to because in most cases the data are either not citable or are made inaccessible via legal and technical hurdles. This is even more tragic in the case of the tropical areas of the global South which are home to the most diversity of species, even as almost all the documented knowledge about this biodiversity resides in the North. Despite technology being available to close this divide, this knowledge remains inaccessible in the South.

For the past 14 years I have been leading Plazi, an organization devoted to building a biodiversity knowledge commons by extracting citable taxonomic treatments and included facts and making them accessible to everyone irrespective of their location.

2. What change do you want to make in the world?

We want to break down the barriers blocking direct access to the results of scientific research, freeing them to be readily available to by anyone, anywhere so their inquiries, discussions, and debates may yield the best and most informed conclusions. We are working on a number of fronts to achieve this objective:

1. We are creating an open scientific knowledge commons through Text and Data Mining (TDM) in place of the highly fragmented and isolated scientific knowledge landscape as it exists now
2. We are developing and promoting semantically enhanced publishing beyond the current PDF and html publishing paradigm, allowing citations of sub-articles and other data elements
3. We are changing the way data are accessed by building an exemplar massive storage of biodiversity related data that will serve as a reference system for knowledge about the world's species
4. We are removing legal hurdles to accessing scientific facts including scientific illustrations that should that should not be encumbered by copyright, and making them citable

5. We are breaking down the domain-specific silos by contributing to the Linked Open Data Cloud
6. We are pushing the global players in the information world such as Wikidata, CITES, IUCN RedList, and NCBI to start linking to taxonomic data instead of just meta- or summary-data

3. What do you believe has prevented this change to date?

Tradition, ignorance, inertia, misreading and misapplication of intellectual property right and commercial interests are the main stumbling blocks for change. We scientists do not realize that our communications are de facto decoupled from the underlying scientific research results. Traditionally, Science has been communicated through articles, and credit has been obtained via citation metrics. The World Wide Web helped to broaden and accelerate access to scientific publications. However, these articles remained semantically unstructured, and the abstracting services are based on metadata of the articles at best. Strong commercial interests – the STM publishers being one of the most lucrative billion dollar businesses – the reuse of the articles for TDM cannot reach the critical mass necessary to change the traditional publishing paradigm. A highly publicized hunt for copyright infringers keeps scientists and administrators from doing anything that could create a legal conflict with the publishers. This inaction is further promoted by misreading copyright law as protecting data.

A main focus of TDM has been to build tools and less so content, such as large corpora of readily citable and minable facts that we consider a game changer.

4. What are you going to do to get there?

At Plazi we have aimed over the last 14 years to lead by example, nimbly and opportunistically developing our own open source tools and services to provide access to a critical mass of high quality, open, and linked scientific data liberated from the confines of traditional publishing. We now aim to sustain our achievement and expand our reach. For that we need to do the following:

1. Install reliable infrastructure and public facing applications and services
2. Establish and document flexible and efficient processing workflows
3. Expand our capacity through recruitment and training
4. Enlarge our reach through active engagement and collaboration with the broad range of creators, publishers, and users of data
5. Continue to develop and enhance our tools and services
6. Provide access to over 50% of the newly described species in the world each year and harvest and process their legacy publications
7. Provide access to and make citable all the extracted illustrations
8. Build awareness about the legal aspects, lobby policymakers for Open Access

9. Maintain and develop a proactive social media presence to build mindshare

5. What challenges or uncertainties do you expect to face?

Evolutionary Challenge: Our approach is novel and disruptive to current publishing practices, it questions deeply rooted behavior of scientists, and as a public service depends on government or philanthropic funding. Ultimately, our hope is that scientific publishing will include machine readable content that does not need prior processing. When that happens, a major part of the Plazi workflow will become irrelevant.

Financial Pressure: In the meantime, the structure of Plazi as a mixture of voluntary and paid work will be put under pressure with increasing production and services. While the core team will last, increasing production and employees will need external support. We are trying to branch into a service organization so we can generate revenue, however, success in that is not assured, and we continue to be dependent on public funding.

Legal Issues: Our legal analysis has concluded that while scientific publications might be copyrighted, individual facts, including scientific illustrations therein, are not. By uploading hundred thousands of illustrations to BLR, we are risking legal action from litigious publishers who may disagree with our analysis.

Technological Capacity: We need to attract skilled collaborators and build innovative application and data flows with innovative partners. For this we need now to bring in more programmers and improve our technical capacity.

6. What part does openness play in your idea?

Openness is fundamental to every aspect of our work. To make the world a better place by building a global knowledge graph depends on linked open data. To us, open data means that the data may be used by anyone without obtaining any further consent, with no restrictions on the outcome whatsoever.

Linked Open Data is an inspiration to us for it not only allows exploring existing knowledge, but also adding additional pieces. We are building a global knowledge graph from millions of mini-silos, the unconnected scientific articles that would allow the community to access, search, and analyse data on their own without needing anyone's permission.

Data are the raw material of science. In our vision, every aspect of the scientific information lifecycle, from its raw material to the tools required to analyze it to the results produced from it are open, linked and citable. We are not just envisioning this future but already have created significant parts of it and are practicing it. We need help to strengthen and complete the lifecycle to realize our vision.

7. Does your idea/project have a name?

TreatmentBank

8. Have you started implementation of the idea?

Yes

9. How have you funded your initiative in the past?

Self funded

Family and friends

Other:

The idea to provide open access to biodiversity literature has originally been funded through a binational US NSF and German DFG fund. Additional funds have been provided through EU research Framework programs (pro-iBiosphere and EU BON) that allowed developing the existing workflow and infrastructure, and pay for meetings and participation in workshops.

A substantial contribution has been through Plazi members. A lot of highly specialize work is done through voluntary input whereby the members can develop and implement cutting edge technologies, data policies, building Impacts that are close to their view of a shared, open world. This reward, the possibility to work on cutting edge, stimulating projects complementary to a salaried job and exposure to a wide network, are a major driver over the last 14 years.

Another input is through in-kind contributions from the National Institutes of Health or Pensoft.

Part of the infrastructure is supported by University of Massachusetts and Zenodo/CERN.

10. Who are your current or potential key partners?

The following ongoing collaboration include partners with global outreach:

Pensoft: Publisher: Collaboration in developing and implementation of subarticle level citation; implementation of JATS/Taxpub to publishing biodiversity literature. Building the Linked Open Data based Open Biodiversity Knowledge Management System (OBKMS).

Global Biodiversity Information Facility (GBIF). Prime user of subarticle elements; global outreach.

Wikidata. Existing collaboration (P1992), automatic input of new scientific facts extracted from new and old scientific articles, and creating necessary entities for the biodiversity domain (eg Plazi Treatment (Q27567957).

Zenodo: Co-developing the Biodiversity Literature Repository, a collaboration to provide access to images, legacy publications, including minting persistent identifiers and metadata. Showcase for providing scientific illustrations as open access. 500K-1M images extracted from literature.

National Center for Biotechnology Information: Developing domain specific Journal Archival and Interchange Tag Suites (Taxpub) for biodiversity literature. TreatmentBank is a target for NCBI taxonomic output.

Potential key partners: Swiss Institute for Bioinformatics: building long term infrastructure

Zazuko: Building Linked Open Data infrastructure for biodiversity data as part of Swiss government long term support and contribution to global infrastructure.

11. Do you intend to implement the idea as a for profit or not for profit initiative?

Not for profit.

12. Where will you be based during the fellowship?

Base Country: Switzerland

Base City: Bern

13. Where will you implement your idea?

Same as above

14. Do you have an online presence?

Plazi: http://plazi.org

TreatmentBank: http://treatmentBank.org

Biodiversity Literature Repository: http://biolitrepo.org

Bouchout Declaration on Open Biodiversiy Knowledge Management: http://bouchoutdeclaration.org

Plazi on Wikidata: https://www.wikidata.org/wiki/Q7203726Applicant's

Twitter TreatmentBank new taxon alert: https://twitter.com/plazi_treat

Twitter account: https://twitter.com/myrmoteras

The open acces version of the curent proposal: https://doi.org/10.3897/rio.2.e11014

15. Does the idea/project have an online presence?

http://treatmentbank.org

http://biolitrepo.org

Link to your video

https://vimeo.com/channels/1144672/185657429

I acknowledge that:

This video is purpose made for this application.

This video link points to a video available on a video hosting service like Vimeo or YouTube.

Curriculum Vitae

Suppl. material 1

16. Have you applied for a Shuttleworth Foundation Fellowship in a previous round?

No

17. How did you hear about the Shuttleworth Foundation Fellowship Program?

Friend

Current/Past Fellow

I acknowledge that

I have read and understood the Foundation Privacy Policy.

I have read and understood the Terms and Conditions of the website.

I am above the age of majority in my country of residence and can legally enter into contracts.

I understand that the Foundation will require all intellectual property created as part of the Fellowship to be openly licensed.

I understand that if my project has a for-profit element, the Foundation will ask for an equity stake based on the investment made.

I understand that if my application is incomplete in any way, it will not be reviewed. It will be excluded for this round.

Acknowledgements

Thanks to Puneet Kishor, Terry Catapano, Daniel Mietchen, Lyubomir Penev and the many un-named olleagues that helped to finally shape this proposal. Thanks to Pensoft and its editorial team to support making this an open proposal.

Supplementary material

Suppl. material 1: Curriculum Vitae

Authors: Donat Agosti
Data type: Text file
Filename: DonatAgostiCurriculumVitaeShuttleworth_20161101.pdf - Download file (402.00 kb)