The Open Information Access Initiative
Working Draft

September 26, 2012

Abstract

This document is a draft of the future policy of the Open Information Access Initiative portal, and more generally of open initiatives for research communities. The OIAI is centered around open journal - the "Open Information Access Journal" - and aim at improving reproducibility and good experimental practices in IA.

Authors

B. Piwowarski, with some comments and ideas from H. Zaragoza, G. Dupret, and R. Blanco.

Goals

This document aims at providing guidelines for the future development of the OIAI portal. Everything is open to debate, including the title!

1 Organization

In order to interact with and to manage the OIAI portal, I think that we would need 4 groups of users:

1. Anonymous readers - anybody can access public material on the website (open papers, discussion, etc.)
2. Members: people allowed to submit papers, participate to discussions, etc. To participate (submission, discussion), people have to be registered. Registration should be reserved to people already contributing to the Information Retrieval/Access community. A system of co-optation would allow new people to join in as readers.
3. Reviewers: people allowed to review papers. I am not sure what would be needed to qualify people as "reviewers". As a start, requiring that people have a research position (university, company). Inviting already existing reviewers from conferences, journals could also be an option. Co-optation (from other reviewers) is also possible.
4. Board: people that can take decisions about the development of the portal and when conflicts arise about reviewers, papers, or any material published on the portal. The board would also ensure the promotion of people between from “anonymous” up to “reviewer”. Board members must have an official position.

Other roles might be needed, but it would be good to limit the number of those to a maximum.

1.1 Voting

In order to ensure the representativity, the board should be elected somehow (maybe not in the first year or two). It is important that there is as much transparency in all the processes and that there is no subgroup that could take too much power - this is vital for the long term run.

I would also like that there is no "head", but that decision can be validated by any group of board members (random). This could be done by:

- Requiring a certain number of board members to validate a decision (number and/or percentage);
- Making all decisions public.

Alternatives to control that there is no group of too closely related board members taking decisions are welcomed! For example, ensuring that the sample is representative enough:

1. Different countries and institution
2. No co-authorship (or not much) or same advisor
3. etc.

1.2 Money

In order to get some fundings (after the initial one) to run the server(s) and pay for additional developments and maintenance, there are a few possibilities:

- Ask for a very small yearly contribution of each member (or organization);
- Ask for (supra)national recurrent fundings.

2 The open journal

I think the portal should accept **any** publication from its registered members right from the start, where papers can get promoted by various ways:

- [repository, arXiv like] by reader-sourcing, using S. Mizzaro proposal as a starting point.
- [light reviewing, PLOS-1 like[1]] by reviewing ("endorsement"): Authors can ask a paper to be reviewed;
- by the editorial board

Papers would have different statuses (not mutually exclusive):

1. Proposal/ideas: kind of "I want to work on that but need help"
2. State of the art.
3. Standard (the typical paper: motivation-related works-theory-experiments-conclusion)
4. Validation papers: I have reproduced the model described in this paper and found those results
5. Source code (i.e. kind of enhanced read-me file)
6. Datasets (description + data)

There could be an update policy that specify what happens if a reviewed paper gets updated. I would say that it is merely linked to the original paper, along with comments by the author(s) and can be submitted again if wanted (as any other paper).

[1] From PLOS website: *Fast, efficient, and economical, publishing peer-reviewed research in all areas of science and medicine. The peer review process does not judge the importance of the work, rather focuses on whether the work is done to high scientific and ethical standards and is appropriately described, and that the data support the conclusions. Combining tools for commentary and rating, PLOS ONE is also a unique forum for community discussion and assessment of articles.*

2.1 Archive

Reader-sourcing (proposed by S. Mizzaro http://www.readersourcing.org): automatic scoring system where readers give a score to each paper. To put is shortly, the closer the score to the mean, the better the rating of the reader.

2.2 Open Journal

The reviewing process should be only loosely connected to the board members.

Each reviewer would be associated to some IR area(s) of expertise[2] and to a set of people or institutions with whom this person has conflicts of interests. This would be a public information.

When a paper is submitted, any (meta-)reviewer associated with the paper main areas could ask to review the paper. Reviewing should be very fast (1 week) so that accepting a paper for review means "I can review it now".

The reviewing process would be lighter than that of a journal, i.e. it should be more a safety check:

- Is the information correct?
- Could the experiments be reproduced?

Papers with innovative ideas could be submitted for promotion to the open journal.

There would be a pool of reviewers and meta-reviewers (taking the decision) that would be associated to IR areas and that could select a paper to review. The review process should be very fast (e.g. a reviewer has one week) and light (i.e. the check is for correctness). There are only three outcomes: conditional promotion, promote or not. In case of "conditional" the set of comments to address should be stated clearly by the meta-reviewer that would ensure that changes are made (maybe by asking the original reviewer(s) to check).

Some part of the reviews could be public - and even not anonymous?

In order to give some incentive to review, reviewing could be done using a system of credits (virtual currency) where you get have to pay in order to be reviewed. Examples:

- Reviewing one paper gives 1 credit;
- Getting reviewed costs 4 credits (3 reviewers + one meta);
- Being late in reviewing costs one credit;
- Maybe some other responsibilities can associated to credit?

The editorial board, if asked by the reviewers, can promote a paper into the electronic journal and be published in a special form - an electronic issue (starting with a quarterly edition?). Special editions could also be edited by members (from reviewer or board members).

2.3 External promotion

A great thing would be to promote to an existing conference (see e.g PVLDB that works like a journal where some papers can be presented at the conference). Maybe we could start by having a few posters in each big conference (SIGIR, CIKM, WSDM, ECIR, ICTIR, OAIR).

[2]The ACM 2012 classification system could be used (provided it is freely usable). See `http://www.acm.org/about/class/2012`

3 Reproducibility and openness

In order to increase the reproducibility of papers, and to allow people to discuss, the portal would

- Use a placeholder for any paper from main publishers (ACM, Springer, etc.) or public archives (arXiv, etc.)
- Associate each paper to its authors that can thus participate to the discussion about the paper
- Link papers (and more generally resources) between themselves so that it is easy to browse the repository of papers. I think the type of links should be restricted if one wants to do something easily with it:
  - extension of a work
  - validation
  - contradiction
  - alternative
  - implements
  - use dataset
- Resources can be associated to papers (code, links, etc.)

The goal is that papers can be discussed at length on the website, comments and remarks being archived for future use.

3.1 Enhancing reproducibility

In order to enhance reproducibility, it would be good to standardize (somehow) experimental practices, and to have an easy way to see what's implemented and, if possible, to run experiments in a short amount of time in order to

Source repository As for papers, version repositories (svn, git) of code corresponding to papers could be open. Comments and updates (especially with git) can be easily shared.

Common standards The idea here would be to describe how the processed (or unprocessed, e.g. documents) information is presented so that chains can be easily implemented (if everybody uses the same format, or set of formats). Converters between formats would ease bridging the gaps.

Experimental manager It is possible to describe experimental components and experiments using an experiment manager that automates experiments, e.g. experimaestro - http://experimaestro.sourceforge.net/. An experimental plan can for example be collection.id=trec.1/adhoc,trec.2/adhoc * model=BM25 * model.k1=0.8,1,1.2

Databases Have an easy way to describe standard (e.g. TREC-like) collections when possible. For example, ir.collections, https://github.com/bpiwowar/ircollections, describes ad-hoc collections and automates some processing (building up the list of files, downloading topics and assessments).

An alternative is to directly report past results. At least two websites/pages report results of common algorithms on standard databases:

- IREval (Sistematic evaluation sistems on TREC collections): http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/
- http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=baselines

I guess a wiki would be a good choice here in order to have the information easily accessible. This is the main bottleneck when trying to compare to previous models.

4 Misc

Finally data should be open as much as possible: people should be able to access easily (and download) all their data in order to lock people in a system (e.g. all the researcher community website that appear). This would be enforced by the definition of an interchange format for any content published on the website (publications, comments, reviews).

The data also has to be stored in a distributed manner (network of university servers?) so that we don't loose any data.

5 Links