Future Professional Communication in Astronomy
(Eds. A. Heck & L. Houziaux, Mém. Acad. Royale Belgique, 2007)

PLETHORIC PROSE VS. SALIENT POINTS

RUDOLF ALBRECHT
ST-ECF/ESA/ESO
Karl-Schwarzschild-Straße 2
D-85748 Garching, Germany
ralbrech@eso.org

Abstract. This paper examines the process of doing research and derives requirements for the interchange of scientific information. These requirements are being mapped into existing and soon-to-be available technology. By way of extrapolation possible improvements to the efficiency and the thoroughness of the research process are identified.

1. Introduction

Communication in science is an integral part of the process of "Doing Science". Thus, before we try to establish the science communication requirements and from these to define the communication processes, it behooves us to try to understand just what "Doing Science" means. Far too many scientist are engaged in this activity without being fully aware of what exactly they are doing.

The model of the research process as developed by Sir Karl Popper comes closest to what most natural scientists do when they "do science" (Popper 1934, 1972). The process consists of several steps:

- Data acquisition, which is the input of signals, either through sensory perception, or through measuring devices which register signals which are either too faint or which are not suited for our senses.
- Transformation of the input data into meaningful values, quite often literally the "data *reduction*" from a jumble of instrument dependent individual readings to a much smaller, coherent and consistent set of parameters.
- By injecting concepts into the collection of parameters we construct models. Concepts range from very simple, such as a linear correlation,

to the very complex, like the hot Big Bang. The injection of concepts happens spontaneously and associatively, it is a genetic disposition and the result of the evolution of our brains.

Models come in two flavors, hypotheses and theories, the difference being that a hypothesis is an as-of-yet unsubstantiated and incomplete theory. Given the fact that no theory is ever complete it is more correct to say that all models are hypotheses. This is in agreement with the historical observation that even "wrong" models served well as good hypotheses in a heuristic sense.

Good models allow to make predictions as to future observations. They also allow to add to our pool of concepts by abstraction and generalization. If a model conflicts with observations we have to discard it. Since we can never be certain that any model will forever withstand the test of future observations Popper concludes that in science we can never demonstrably attain the "truth".

In actual fact the research process is just the extension and extrapolation of the mechanism which humans have used in the past to cope with their environment, and as such it is a genetic disposition. Asking the question where in this process the most progress has been made historically, we tend to think that it has been in the first step: the introduction of ever more powerful telescopes and detectors, and the opening of more spectral windows has allowed to quite literally include observations of much of the universe into the building of models. I should contend, however, that the most significant progress has been made in the application of concepts: the scientific revolution (i.e. paradigm change) during the period of enlightenment removed concepts like that of the supernatural, of magic and of subjective notions from our model building tools, which indeed provided us with the very basis of what we today call scientific thinking.

There is a caveat: tempting as it might be to think that we have found the key to understanding nature, it is neither evident nor indeed likely that the scientific methodology we use today is capable of providing us with a full explanation of the universe. With our limited access to the universe (essentially only the electromagnetic signatures), and probable severe limitations to our concept generation abilities we are unlikely to achieve the desired result (Albrecht 2000a).

2. Evolution of the Methodology

Modern science began with a generation of "deep thinkers" (Leonardo da Vinci, Tycho, Newton, Kepler, Galileo), who were able to collect all relevant knowledge available at the time in their brains and, after years of pondering, published great tomes (e. g. "De revolutionibus ...") , which represented the

state-of-the-art for several decades. Quite often the deciphering and the translation (from Latin) took more intellectual effort than understanding the content. This approach prevailed well into the 19^{th} century.

We have come to realize that this process is not efficient: sitting on information for decades is unsatisfactory, and individual human brains are no longer capable of holding all required information. Science has become a collaborative effort, and good communication is thus a prime requirement. Communication ("papers") are also important as proof of ownership, as the output product, and to justify funding. Communication (in the sense of public information) has also become important to advertise and raise funds.

3. Science Communication

Notwithstanding the fact that media and distribution mechanisms have changed the basic principle of science communication is still Gutenbergs technology: the human author uses a natural language enriched by technical terms and structured by agreements and conventions to commit the information to a medium, hoping that readers will be capable of re-enacting the cognitive process of the respective author in their brains.

Given the heterogeneous educational and cultural background, different native languages, different writing style and ability, different gender, ages and social context of the readers, this is not a reliable process.

In astronomy we have at least successfully standardized on a common language: scientific English. This provides a basic level of readability, but does not guarantee universal understanding. Possible solutions could be the definition of a define a meta-language.

Compared to scientific papers of several decades ago, today papers have become more narrow, more terse, more solid, and much more numerous. While the delay between submission and publication has come down, printed papers have assumed a "for the record" type of role. With the advent of the web the network has become the prime distribution medium and is used, almost exclusively, by the working scientist. The new medium offers functionality which the printed version can't (access to original data, access to literature references, forward references, visualizations, links to related or supporting information, computing capabilities, etc.). This has become known as the "3-D Publication model".

In fact, the available technological possibilities are so plentiful that they have not been fully utilized yet.

4. The New Media

It is becoming evident that around the turn of the millennium the traditional scientific library started to become anachronistic for a number of reasons. For the librarian a publication needs to be a final product which henceforth will neither change nor decay. (Ideally: stone tablets). For the working scientist on the other hand the paper should reflect the latest stage in the author's research, if possible with dynamic updates, which is the exact opposite. In combination with long publication lead times this means that by the time a paper appears in the library it is already obsolete.

The electronic media can remedy this situation. However, keeping this process under control requires a complete change of the publication paradigm. Major disadvantage: It is, at this point, totally impossible to perform computer-assisted logical operations on a paper or a set of papers, such as discovering supporting or contradicting evidence, comparing the paper with a model, injecting hypotheses into a collection of papers and check for consistency.

5. Issues and Problems

The smallest unit of information in astronomy is still "the paper". It depends on the literary skill of the author how well the paper conveys the information, and it depends on the selection of title, keywords and other ancillary information how reliably the paper can be recognized as relevant by another scientist.

Even though powerful tools have become available, like the Astrophysical Data System (Eichhorn et al. 1994), it is difficult to identify AND READ the relevant papers in subdisciplines of astronomy other than one's own. It is impossible to identify AND READ possibly relevant papers in other related disciplines (e.g. nuclear physics). Even though full text analysis is technically feasible, it suffers from the ambiguities of encoding content in a natural language ("planetary nebula" vs "planetary atmosphere", stellar magnitude vs. order of magnitude etc., etc.)

The "paper" is, for historical reasons, a holy cow. However, what the working scientist really needs is not "the paper", but relevant INFORMATION.

Attempts to use neural network techniques (Kohonen 1995, Lagus et al. 1996) for full text analysis in order to identify mutually related papers and to group them according to relevance have been made (Lesteven et al. 1996, Albrecht & Merkl 1998). Although the results were promising the approach has not been implemented on a scale which might make it useful.

An example can be viewed online[1] (Albrecht 2000b).

6. A Historical Analogy

Ever since mathematics was started to be used in astronomy the numerical operations were performed at the limit of the human brain: human "computers" became the bottlenecks. The next step was to assist the human brain by using tools like logarithms and finally mechanical devices. Still, the human operator had to lay hands on each and every number.

About 50 years ago, the requirements for numerical processing started to exceed the ability to look at each and every number. Automatic (programmed) processing started. Initially there was considerable resistance towards "trusting a machine", but in the end, the obvious benefits were convincing (stellar evolution calculations, orbit analysis, Monte-Carlo experiments)

I contend that we are in a similar situation when it comes to combining information from different sources: the processing requirements are beginning to exceed the capabilities of the human brain, both in terms of throughput and in terms of complexity.

7. Back to "Doing Science"

Scientists do not start on a project by deductively deriving a model (concept, hypothesis, theory) from the input data. Rather, they approach the input data with a model in mind, which they will alter/refine during the analysis ("school of thought, bandwagon, prevailing thinking").

Scientific breakthroughs are achieved by finding new models (Kopernikus vs. Kepler), or by realizing that several seemingly disassociated phenomena can be explained by one and the same model (Newton).

Inventing and applying models is, at this point, still the domain of the human brain. Similar to the evolution in numerical processing we have developed methodologies to help the brain in the process: morphology, taxonomy, classification, etc. More recently: visualization.

The human brain is a bottleneck, also for reasons of principle: it evolved on this planet and served us well to survive on it. It is far from evident that this made it capable of generating proper models of the universe at large. We can circumvent this problem by dropping the requirement of practical understanding and relying solely on the logic of mathematics (e.g. superstrings). In full analogy to numerical processing, where we trust the computer with calculations which humans could not do, and only look at

[1] `http://www.stecf.org/~ralbrech/papers/aheck/kmap3.html`

the result, we have to enlist the help of automatic processing to generate and verify/disprove models and hypotheses.

8. Proposal

We should stop distributing information as "papers", but instead (or, initially, in addition) as updates to a computer-processable data base (knowledge base). This would require the definition of a dictionary/thesaurus of terms and operators, implemented in a language like XML.

For a pilot project we should identify a suitable (i.e. small enough) subfield of astronomy (TNOs?) and start to build up a fully processable data base. The original papers could still be there, as comments and appendices. Ideally we should convert and include the relevant papers of the last 20 years. The best initial pool of such papers would be the "Letters" series of the major journals

Data mining would immediately yield implied knowledge, which we do not even know we have. After initial experience with such a pilot project we should approach other disciplines to consider this approach, in particular high energy physics.

9. Can we do This?

Not within astronomy. We need the help of computer science and we need significant funds.

Considerable related work is being done in the area of natural language/speech analysis (translation, homeland security). The common denominator is that garbled and otherwise disorganized and even deliberately distorted input data have to be inserted in a process of associations to build hypotheses.

Signal processing and image analysis is another example. It takes input data and uses a sequence of segmentation and classification to identify patterns and items in the image, and takes action according to their nature (computer vision, airport security).

Web browsing is the prime example of multi-faceted, partly overlapping and quite often inconsistent data sets, with a volume which by far exceeds the capability of the human brain to process it within a reasonable period of time. Considerable research has been conducted (Dittenbach et al. 2006). So far the emphasis has been on speed and completeness (Google). Efforts are being made to achieve best relevance (e.g. Mooter).

Content processing requires similar capabilities. It is being persued aggressively, as it is important to maintain computer network security.

The process of information complexity is well known: the US National Security Agency apparently had all the evidence to predict the assassination of Anwar Sadat. They just did not know that they had it.

10. Steps we can Take

It is evident that we cannot change our publication methodology from essentially colloquial English plus conventions to something computer processable on a short timescale, but we can do first steps:

- Publishers could require a structured abstract of the type: hypothesis-assertion-evidence-conclusion, or similar (Bertoud & Schneider 2005). Templates could be provided.
- We can collect these abstract institutionally (large observatories, VO, ADS, Simbad) in a suitable distributed data base (archive) and turn them into XML using automated tools.
- The input language would initially be English, but this is not necessarily a requirement.
- We can provide VO-type tools to view, analyze and process such abstracts.

11. Consequences and Advantages

New information which is to be added to such a data base could immediately be checked against existing information. Agreement or disagreement can be identified and can be brought to the attention of the scientist(s).

Plagiarism or already known science would be instantly recognizable. Authors would in fact not submit such information for publication. This would also provide the ultimate objective way to measure productivity and impact.

Refereeing becomes trivial, as "new" science can be unambiguously identified. Curator(s) would be needed to perform such operations and to maintain the integrity of the data base and the associated tools.

New information does not necessarily have to be based on observations. It can also be generated in a what-if type of manner. Scenarios can be processed Monte-Carlo style. We could even postulate far-fetched models like a time- or distance dependence of the laws of physics and check for possible consistency.

The impact of new information would be extremely objective, i.e. even an unknown graduate student, or an astronomer working in a developing country would be instantly recognized for a possibly fundamental contribution.

The contents of the data base (or any subset thereof) can be mapped into a (in fact, into any) natural language by using appropriate translation tools. The natural language representation can be at different levels of by using a thesaurus to expand the information, or to different depth.(i.e. professional communication and public information; "customized journal").

The information could also be turned into a Wikipedia-type of representation, with the advantage that it would be based on verified information, and not on the personal knowledge of individuals.

Once this method catches on it would be possible to combine more subfields of astronomy, and, in the long run, even interdisciplinary fields.

Finally, and most importantly, the tedious process of postulating, trying and re-trying, refining and iterating would be done, tirelessly and efficiently, by a machine.

References

1. Albrecht, R., 2000a, On Possible Limits to Understanding the Universe, in *The Future of the Universe and the Future of our Civilization*, Eds. V. Burdyuzha & G. Khozin, World Scientific Publ., Singapore, p. 53.
2. Albrecht, R., 2000b, Computer-assisted context analysis of data bases containing scientific literature, in *Information Handling in Astronomy*, Ed. A. Heck, Kluwer, Dortrecht, p. 109.
3. Albrecht, R. & Merkl, D., 1998, Knowledge Discovery in Literature Data Bases, in *Library and Information Services in Astronomy III*, Eds. U. Grothkopf, H. Andernach, S. Stevens-Rayburn & M. Gomez, Astron. Soc Pacific Conf. Series **153**, p. 93.
4. Bertoud, C. & Schneider, P., 2005, Introducing structured abstracts for A&A articles, *Astron. astrophys.* **441**, E3.
5. Dittenbach, M., Berger, H. & Merkl, D. 2006, Automated concept discovery from Web resources, in *Proc. IEEE/WIC/ACM Int'l Conf. on Web Intelligence* (WI 2006), Hong Kong, IEEE CS Press.
6. Eichhorn, G., Murrey, S.S., Kurtz, M.J., Accomazzi, A. & Grant, C.S., 1994, The New Astrophysics Data System, in: *Astronomical Data Analysis Software and Systems IV*, Eds. R.A. Shaw, H.E. Payne, J.J.E. Hayes, Astron. Soc. Pacific Conf. Series **77**, p. 28.
7. Kohonen, T., 1995, Self-organizing maps, Springer-Verlag, Berlin.
8. Lagus, K., Honkela, T., Kaski, S. & Kohonen, T., 1996, Self-organizing maps of document collections – A new approach to interactive exploration, in: *Proc. Int'l Conf. on Knowledge Discovery and Data Mining*, Portland, OR, p. 238.
9. Lesteven et al., 1996, Neural Networks and Information Extraction in Astronomical Information Retrieval, *Vistas in Astron.* **40**, 395.
10. Popper, K., 1934, Die Logik der Forschung, Springer Verlag, Wien.
11. Popper, K., 1972, Objective Knowledge, Oxford Univ. Press.

Links

`http://www.stecf.org/~ralbrech/papers/aheck/kmap3.html`
`http://www.eric.ed.gov/ERICWebPortal/resources/html/help/help_popup_submission_structured_abstract.html`

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=128954
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=116403
http://ausweb.scu.edu.au/aw2k/papers/lowe/paper.html
http://www.mooter.com/
http://www.msdewey.com/