The Open Access interviews: Annette Holtkamp

Richard Poynder talks to Annette Holtkamp, an information professional at Germany's largest particle physics research centre Deutsches Elektronen Synchrotron (DESY). Holtkamp is a member of the Helmholtz Society Open Access working group, and the German contact for SCOAP³. More recently she has been heavily involved in the development of a new repository for the particle physics community called INSPIRE.

Physicists are widely credited with having begun the Open Access (OA) movement, on the basis that in 1991 theoretical physicist Paul Ginsparg created the physics preprint repository arXiv — an Internet-based service that allows physicists to electronically share their preprints with one another prior to publication in a scholarly journal.

In reality, however, the seeds of the OA movement predate arXiv by at least thirty years, when scientists working in the specific domain of particle physics — otherwise known as high energy physics, or HEP — began to routinely share their preprints via the postal system.

Of course scientists have always exchanged papers with one another by mailing copies of their papers to friends and colleagues. What was new, however, was that during the 1960s librarians at major HEP research institutions like DESY, the US-based Stanford Linear Accelerator Center (SLAC), and the world's largest particle physics laboratory CERN, began developing bibliographic tools and services to assist scientists share their preprints in a formulised fashion, thereby allowing researchers to exchange papers in a new and more effective way.

The service that came to dominate was SPIRES, which began life as a preprints collection and an associated card catalogue system created in the SLAC library in 1962. In 1974 this was computerised and combined with work being done at DESY, broadening the service to include bibliographic data on journal articles, conference proceedings, theses and books. Importantly, a number of SDI services were also introduced. These allowed HEP scientists to request that they be alerted when new papers in their field became available. Since the alerts included author contact details, scientists were able to communicate directly with one another in order to exchange papers, regardless of whether they personally knew each other.

When the Internet became widely available new features were added to SPIRES so that researchers could interrogate the database by email, allowing them to retrieve the information electronically. And in 1991, SPIRES became the first web-based database.

SPIRES has always been a bibliographic database. What was innovative about arXiv, therefore, was that it was the first freely available electronic repository containing the full-text of research papers. However, since SPIRES offered powerful search features, citation analysis tools, and was actively managed by librarians (who started inserting links into its records to allow users to go directly to the full-text on arXiv), researchers began to use SPIRES as a search engine for arXiv. As such, a symbiotic relationship developed between the two services.

Today SPIRES has around 770,000 records and is growing at around 4,000 new records a month. ArXiv, meanwhile, has 492,000 e-prints, with roughly five thousand new ones added each month.

But SPIRES is now over thirty years old, and had begun to creak. Consequently, in May a new service called INSPIRE was launched. INSPIRE is a joint initiative of SLAC, DESY, the US-based Fermi National Accelerator Laboratory (FNAL) and CERN.

The new database is based on a state-of-the-art Open Source software platform called Invenio, initially developed for CERN's institutional repository — known as the CERN Document Server, or CDS. It is planned to populate INSPIRE with content from both SPIRES and CDS — which has around one million records, half of which are full text — promising to make INSPIRE a significant new service for the HEP community.

What is the long-term objective for INSPIRE, and what does its development portend? An information professional at DESY for 25 years, Annette Holtkamp is ideally placed to answer these questions. Holtkamp has long been responsible for inputting journal and conference proceedings

into SPIRES, and is directly involved in the development of INSPIRE. "Our ambition is to build a comprehensive HEP information platform hosting the entire body of HEP metadata and the full text of all OA publications", she says.

As such, INSPIRE will eventually become a full-text service like arXiv — something that in a recent survey HEP scientists said they wanted. Moreover, it will build on the strengths of SPIRES. To this end it was decided to use very sophisticated repository software which will, amongst other things, include a new search engine, a metrics system for measuring the impact of articles, and a number of specialist data management tools. It will also boast Web 2.0 functionality and, as with SPIRES, the data will be managed and curated by professional librarians.

What implications might this have for arXiv? Can we expect particle physicists to discontinue putting their papers in arXiv, for instance, and opt for INSPIRE instead? It would, after all, seem a more natural location for HEP researchers to deposit their preprints. Not necessarily, says Holtkamp. "HEP scientists have already a full commitment to self-archive their preprints in arXiv and there is no reason to change this. What INSPIRE will offer is the possibility to deposit material that users want to be preserved in addition to today's possibilities, such as theses or old preprints which were not submitted to the arXiv originally: indeed, authors are usually happy to make them available but arXiv is seldom used as such a forum."

INSPIRE needs also to be seen in the context of the wider OA movement. Unsurprisingly Holtkamp is a passionate advocate for Open Access. She is a member of the Open Access working group of the Helmholtz Society, and was on the working party that designed SCOAP^3. And when Germany joined the SCOAP3 consortium she became the German contact for SCOAP3.

SCOAP^3 is an ambitious project that hopes to "flip" the entire particle physics literature from today's primarily subscription-based model — in which researchers (or their institutions) pay to access published research — to an Open Access model, in which the HEP research community would instead pay to publish its research. (Not on an author-pays model, but through a single consortium that will facilitate the re-direction of subscription funds). In return, publishers would commit to making HEP papers freely available on the Web.

The development of INSPIRE, however, opens up the possibility that SCOAP^3 could prove to be a transitory phase. If successful SCOAP^3 would of course make all HEP research OA. But as a result of the so-called serials crisis — and growing disillusionment with the role that publishers play in scholarly communication — some believe that it is time for the research community to, as they like to express it, "take back ownership" of its research.

In other words, the argument goes, it is time not only for researchers to stop signing over copyright in their papers (something traditional publishers generally insist on as a condition of publication) but for the research community to begin to take responsibility for distributing its own research too. Holtkamp is keen to stress that this is a minority view however. "The bulk of the community wants to preserve the journal system," she insists. "And this is definitely not on the SCOAP3 agenda."

Nevertheless, we should note that one consequence of the rapid rise in both subject-based and institutional repositories is that scholarly communication is moving from a journal model to a database model. This could see a second kind of flip take place: Where today researchers pass over their papers to publishers, who then become the official source and distributors of published research, in the future repositories could become the primary location for scholarly papers. They might also become publishing platforms, with publishers relegated to outsourced service providers paid simply to organise the peer review of the papers that have been deposited in repositories. This, for example, seems to be the model that the University of California is moving towards.

As such, SCOAP^3 and INSPIRE are perhaps just the first signs of a far more radical revolution poised to sweep through scholarly communication; one that was never envisaged when HEP librarians began developing new tools to help particle physicists share their preprints with one another.

One related new development, for instance, is the growing desire — nay necessity — for what is called "Open Data". In contrast to research papers, the issue here isn't just that of ensuring that the

scholarly community retains ownership of its research, nor just that it ensures the data generated by experiments is preserved in a machine-readable format (Bearing in mind, for instance, that with data formats being constantly updated and superseded we currently face the threat of a digital dark age).

As important as these issues are, the main concern for particle physicists today is that in the light of the increasing complexity of the experiments they conduct it isn't enough simply to keep data in its raw form (even if it is constantly migrated to new machine-readable formats), but to retain with it the knowledge necessary for anyone who did not take part in the original experiment to reuse and reinterpret it. As Holtkamp puts it, "We will have to develop what we call a parallel format — that is, a format that not only preserves the data itself, but also the necessary knowledge to be able to interpret it."

What is the nub of the issue here? Simply that in a world of Big Science — where conducting experiments requires building massive particle accelerators like CERN's new Large Hadron Collider (LHC) — it would be a huge waste of money if the data produced could not be reused. It cost 6 billion Euros to build the LHC for instance. If the HEP community allows the 15 petabytes of data the LHC will generate each year to wither on the vine, and rapidly become obsolescent, it would be sheer profligacy.

As CERN director general elect Rolf-Dieter Heuer put it to me in a recent interview for *Computer Weekly*, "Ten or 20 years ago we might have been able to repeat an experiment. They were simpler, cheaper and on a smaller scale. Today that is not the case. So if we need to re-evaluate the data we collect to test a new theory, or adjust it to a new development, we are going to have to be able reuse it. That means we are going to need to save it as open data."

For the moment, however, the particle physics community has no way of doing this. "This is a task for the experimental physicists, working with the IT people, not for library-based information professionals like me," says Holtkamp. "And right now they are just at the beginning of the process. The people conducting the experiments are going to have to sit down with the IT people and work out how to do it."

She adds, "I am pretty confident that Open Access will be the standard of the future for scientific papers, although it remains unclear when Open Data will become the norm."

What is striking is that even though the seeds of the OA movement date back over forty years, the full implications of the changes set in train by HEP scientists are only just beginning to be apparent. As they do, we can surely expect a much broader and far more profound revolution in scholarly communication than anyone could have predicted back then and in all disciplines, not just physics.

In addition to Open Access and Open Data, for example, we are seeing the development of a growing number of open lab notebook initiatives, projects like OpenWetWare, where scientists are able to post and share new innovations in lab techniques; initiatives like the Journal of Visualised Experiments (JoVE), an OA site that hosts videos showing how research teams do their work; GenBank, an online searchable database of DNA sequences; the Science Commons, a non-profit project focused on making research more efficient via the Web (e.g. by enabling easy online ordering of lab materials referenced in journal articles), scientist-friendly social networking sites like Laboratree and Ologeez, and a growing emphasis on the use of Open Source software by scientists.

Collectively this broader and rapidly-developing phenomenon has been dubbed Open Science. As Live Science senior editor Robin Lloyd put it recently, "Open science is a shorthand for technological tools, many of which are Web-based, that help scientists communicate about their findings. At its most radical, the ethos could be described as 'no insider information.' Information available to researchers, as far as possible, is made available to absolutely everyone."

In the "Principles for open science" published on the Science Commons web site four main pillars are seen to open science: Open Access to research literature, Open Data, Open Access to publicly-funded research materials (cell lines, DNA tools, reagents etc.), and an open cyberinfrastructure.

So how might the research process look in this new world? We don't yet know. Holtkamp, however, has some ideas. "I can see a piece of research starting with a researcher simply putting an idea on a wiki, where it would be time-stamped in order to establish precedence," she says. "Others could then elaborate on the idea, or write a program or do some calculations to test it, or maybe visualise the idea in some way. Publication would then consist of aggregating all the pieces of the puzzle — all of which would be independently citable."

This surely assumes a future research environment in which there will be few if any scientific secrets and the traditional model of the scholarly journal will be increasingly marginalised, and perhaps entirely superseded. Above all, it implies that the research community will increasingly expect to retain ownership of its research, and it will take far greater responsibility for communicating and sharing its findings.

Of course it is the increasing complexity of science, and above all the rise of the Internet, that have made these developments inevitable. Certainly they would not have been possible without the Internet. One is nevertheless bound to wonder why the seeds of this revolution first sprouted in the garden of the HEP community, and why so early on? What is it about particle physicists that made them OA pioneers?

With a degree in sociology (as well as a PhD in physics), Holtkamp is better qualified than most to suggest an explanation. That physicists were so early in the game, she says, reveals something about their mentality. "If they encounter a problem they immediately want a solution. If nothing ready-made is available — a very common situation — their strong self-confidence and pronounced playfulness lead them to sit down and try to work it out themselves, and often with success"

It is a mindset that always saw physicists loath to have to wait months (or longer) before reading about a new piece of research. They have always wanted immediate, free access to the latest findings — something that the long lead time and subscription barriers inherent in the traditional journal publishing model could never properly satisfy.

From this perspective, one might want to suggest that it was not that the Internet provided the stimulus for scientists to change the way they did things, but that with its arrival physicists were finally able to share their research in the way they needed to. The Internet simply allowed them to satisfy a long-standing pent-up need?

If you would like to learn more about how physicists sparked the revolution now sweeping through scholarly communication, please read the interview below with Annette Holtkamp.

In the beginning

RP: Can you start by giving me a brief overview of how physicists have shared their research historically?

AH: It varies within the field of physics, but particle physicists have always been dissatisfied — and still are — with the time gap between writing a paper and its publication in a journal, and so developed a strong preprint culture very early on.

It has been standard practice for over 40 years now for HEP scientists to distribute their preprints to one another. Initially this was done by means of the postal system, either by the researchers themselves or by their libraries — who in the early days mailed out hundreds, or even thousands, of paper copies directly to researchers at other institutions, or to other HEP libraries.

RP: This was done without charge was it?

AH: Yes, and it was a substantial financial burden. By the 1990s, for instance, it was costing DESY about DM1 million or 500,000 Euros [$770,000] a year to distribute its researchers' preprints to other researchers and their institutions.

RP: In a sense then one could argue that Open Access has been the norm for particle physicists since the 1960s, although clearly the method of distribution has changed over time.

AH: Indeed, and you could characterise the early postal distribution system as a form of institutionally-paid Open Access.

RP: HEP libraries became involved right from the beginning did they?

AH: Absolutely, and it wasn't simply a matter of putting papers in the mail. They also began collecting preprints, classifying them, and distributing information about them. In the early 1960s, for instance, the libraries at both DESY and SLAC [The Stanford Linear Accelerator Center] began distributing lists of recent accessions.

RP: This was an alerting service?

AH: It was. The DESY library started doing this in 1963, and published a bi-weekly HEP index containing bibliographic information about preprints, as well as journal articles, conference proceedings and so on. These were grouped according to standardised keywords. And we also offered an SDI service.

RP: SDI implies that researchers could specify particular keywords and the library would automatically send them the bibliographic details of any papers matching those keywords as they became available?

AH: And these were distributed to scientists all around the world.

RP: So if a paper matching their keywords became available researchers would contact the library and asked to be mailed a copy?

AH: Actually the preferred method was to contact the authors themselves. We distributed the bibliographic details, but if someone wanted a hard copy they would usually contact the author directly.

RP: Interesting. This means that the author (or the author's institution) paid the distribution costs, so in a sense you could argue that it was a form of "author-pays" OA! The system developed by the SLAC library for managing this process was named SPIRES wasn't it?

AH: Not quite. In 1962 the SLAC library started to collect preprints and to catalogue them. This preprint catalogue later evolved into SPIRES HEP.

RP: SPIRES was originally a card index system right?

AH: The preprint catalogue originally was. In 1969, Stanford University developed the SPIRES (Stanford Public Information Retrieval System) DBMS with the SLAC library being its primary test site. In 1974 the SLAC preprint catalogue turned into the SPIRES HEP database. In the same year, SLAC and DESY joined forces, so that SPIRES contained in addition to preprints also information on journal articles, conference proceedings, theses and books. In 1984 it then became possible to query the database via email. And in 1991 SPIRES became the first Web-based database.

RP: It was also in 1991 that Paul Ginsparg created arXiv, another Web-based physics database. What is different about arXiv, however, is that it hosts the full text of papers, not just bibliographic data. It was also the first self-archiving initiative, with physicists encouraged to post their preprints in a central server, rather than send multiple copies to everyone wanting to read their papers.

AH: Which has made arXiv the perfect complement to SPIRES? ArXiv provides the full text but it doesn't put any emphasis on sophisticated searching or curation of the information. SPIRES, by contrast, is a curated metadata-based service. From the beginning there was a very close partnership between SPIRES and arXiv, and SPIRES has linked to arXiv since 1992. In return we feed publication information to arXiv.

RP: In effect, SPIRES became a search interface for arXiv?

AH: One could say that arXiv is a submission interface to a repository offering the added value of alerts, and SPIRES is the search interface to the entire HEP corpus, both arXiv and published journals. There is a kind of symbiosis between arXiv and SPIRES — to the extent that today there is some confusion in some users' minds over how SPIRES and arXiv actually differ.

INSPIRE

RP: Ok, so SPIRES began as a card index system for HEP research that contained bibliographic information. This later became an electronic database that researchers could interrogate remotely, first by email and then over the Web. Then, when in 1991 arXiv started offering a full-text service, physicists were able to search on the bibliographic data in SPIRES and link directly to the full-text documents in arXiv. And in May this year a new a new database called INSPIRE was launched, which is a joint initiative by CERN, DESY, SLAC and Fermilab [The Fermi National Accelerator Laboratory]. Why does the HEP community need yet another database?

AH: Because none of the current systems is yet able to fulfil all the needs of the HEP community. SPIRES, for instance, suffers from being based on aging technology: the underlying database engine is now over 30 years old, and there is only one maintenance programmer in the world committed to developing it. Moreover, its features have tended to be added in a very ad hoc manner, so the code base can no longer be changed without a massive effort.

In short, SPIRES is suffering a slow paralysis, and users are starting to notice this. Last year, for instance, we sought the views of the HEP community by conducting a user survey. When we looked at the results we found comments about SPIRES that said things like, "The user interface is too arcane" or "The system is too slow."

RP: I believe INSPIRE will be based on the Invenio software platform, which was originally developed for the CERN Document Server (CDS).

AH: Correct. CDS is based on a more modern platform, providing full text access, improved search capabilities — and it's very fast. However, CDS does not offer the same depth of coverage or extent of data curation and enrichment as SPIRES and is therefore less popular with the HEP community as a reference repository. So it seems quite natural to join forces and move SPIRES to Invenio.

RP: CDS is CERN's institutional repository (IR) isn't it?

AH: CDS started in the late 1990's with the purpose of managing scientific information at the laboratory, but now has a double role. It is CERN's IR, but it also expanded into a gateway to HEP information at large, indexing the content of major journals and harvesting full text from many preprint servers, with most of the content coming from arXiv. But these efforts are more limited in scope and time than those at SPIRES.

RP: That is an interesting development. The purpose of an institutional repository, after all, is simply to deposit the work of researchers in a specific institution. CERN has decided to expand the role of its IR?

AH: Given the central role of CERN in HEP this happened sort of naturally, following the needs and expectations of the users. This is something quite central to the way we develop services in HEP

RP: So INSPIRE will be a modern version of SPIRES, but one that combines the content of both SPIRES and CDS in one database?

AH: It will combine the content of SPIRES with that part of CDS that is relevant to the HEP community at large. And there is the additional benefit that it will eliminate the duplicate work that has been going on with CDS and SPIRES, and so free up resources for innovative projects. Our ambition is to build a comprehensive HEP information platform hosting the entire body of HEP metadata and the full text of all OA publications.

RP: Talk me though the content in the SPIRES and CDS databases?

AH: SPIRES has over 770,000 records and is growing at around 4,000 new records a month. As we discussed, this is all metadata.

RP: When you say that all the records in SPIRES are metadata you mean bibliographic data?

AH: Bibliographic data including links to full text, citation analysis tools, links to other SPIRES databases providing information on conferences, institutions, experiments, people, jobs. Meanwhile CDS has almost one million records, about half of which are full text. It also has a growing multimedia collection, including videos and photos.

RP: For purposes of comparison, arXiv currently has around 500,000 full-text papers. But as I understand it the purpose of INSPIRE is not so much to add new content but to combine the content and the features of the SPIRES and CDS databases.

AH: That is our starting point yes. Please note, by the way, that while SPIRES HEP will be completely replaced by INSPIRE, CDS will continue as CERN's institutional repository. But as I said, we will take all the material relevant to the HEP community from CDS and deposit it in INSPIRE. For instance the folks at the CERN library have collected in CDS more than 10,000 theses and have recently put online all the articles ever written in theoretical physics at their lab, this is 11,000 Open Access pre-prints dating back to the 1950's. Along these lines, we plan to extend the scope of

INSPIRE by including older records, other material like conference slides and more articles from neighbouring fields.

RP: At the same time you want to migrate SPIRES to the Invenio platform in order to provide new features and improved functionality? Can you tell me more about the Invenio platform?

AH: Invenio is an Open Source digital library software platform that tries to couple traditional library techniques with modern web technologies. It was developed as proprietary software at CERN in around 2000, but subsequently released as free Software in 2002.

RP: So it would be comparable to repository software like DSpace, but tailored to physics information perhaps?

AH: No, it's not specific to physics. Invenio has been installed in about 20 different institutions and library networks worldwide, and it is being used to host content from a range of different disciplines. What *is* specific about Invenio is that it is targeted at large repositories, from several tens of thousands of records to several millions. It is also designed to host documents of a very diverse nature: preprints, audio, video etc. Key features of Invenio are configurable regular and virtual collection trees, a high-performance native search engine, flexible metadata formats and collaborative features. However, the price to pay for its flexibility and high performance is that you increase the complexity of the system.

RP: Which presumably has management and support issues?

AH: You have to run e.g. several native indexing and pre-caching daemons during runtime. There are, by the way, a few other new things about INSPIRE that I would like to mention.

RP: Please do.

AH: One thing we plan to do is to create much improved tools for citation analysis. This, for instance, will allow researchers to look for papers that are cited together with other papers.

In addition we are thinking of installing a new metrics system for measuring the impact of articles, including articles by individual authors and articles by groups of authors. And we want to develop new ways of differentiating between various ranking systems, including download and citation counts.

Another application could be to offer contextual searching of metadata, for instance to work out if we can suggest a referee if you are a journal editor (or an expert to contact if you are a person seeking advice) by analysing a paper for its citations, where these occur in the paper, who the author has cited before and/or has worked with before and so on.

Finally, we plan to offer a number of Web 2.0 applications.

RP: Can you give me some examples of the Web 2.0 applications you will offer?

AH: Well, one feature already built into Invenio allows people to review, comment, or rank articles — so we will utilise that. Another thing we'd like to develop is a way to present in a single place all comments that all possible sites allowing "living papers" have collected, so that there is an integral view of what a paper has collected. We also have a dream of developing new literature awareness tools like those used by The Faculty of 1000, for instance.

RP: Faculty of 1000 could be described as a post-publication filtering service I guess, where well-regarded researchers are asked to highlight papers they believe to be important, particularly papers that have not been published in a high impact journal?

AH: Yes and another thing we intend offering is subject tagging.

RP: When you say subject tagging you are talking about folksonomies?

AH: Not really. Here at DESY we have been assigning standardised keywords to papers for more than 40 years. These are devised by human indexers.

RP: Information professionals presumably?

AH: Information professionals and active physicists. However, we are currently evolving our thesaurus into a HEP taxonomy which serves as basis for automatically assigning keywords to papers.

RP: So the terms will still be decided by information professionals, but the assignment of them will be automated?

AH: What happens today is that the automatically assigned keywords are corrected and improved by human indexers. But we realise that this process could probably be improved by engaging the community. So we might want to allow authors to change the keywords assigned to their papers, or suggest new terms for the taxonomy in accordance with the development of the field.

RP: In other words, when a researcher deposited a paper they would be able to say, "Ok, I see a number of keywords have been automatically assigned to my paper, but actually I don't think those terms define the topic of my paper adequately. I'd like to suggest a few totally different terms."

AH: And we think that an approach like that will circumvent the known drawbacks of folksonomies. After all, we've spent forty years creating a standardised system; rather than just throw that work away, it seems much better to encourage scientists to build on that work by suggesting new concepts that can be added to it.

RP: As such the taxonomy would retain a traditional top-down approach, but you would invite bottom-up suggestions for improving it?

AH: Yes exactly. Another thing we are working on is a unique author identification system, which I think is very important. We already have a database called "HEPNames" within SPIRES, which contains a lot of information about HEP scientists. The plan is to use this database to uniquely link publications to the author(s). And once again we envisage inviting reader input in order to improve the data.

RP: Clearly HEPNames is specific to particle physics. I'm aware that there are a number of wider initiatives within the research community aimed at creating a unique author identification system for the whole of academia — initiatives like Thomson's ResearcherID and Elsevier's Author ID.

AH: That's true, but developing a HEP specific-system is a simpler process. Nevertheless, it is clearly important that whatever we do is compatible and interoperable with other systems, so we are in discussions with both Elsevier and Thomson.

Another thing we plan to do in the future is to aggregate all related objects.

RP: How do you mean?

AH: One feature respondents to our survey said they would like, for instance, is the ability to group together in their search results all related items, including articles, preprints, conference slides, software representations, videos etc.

RP: Allowing them, for instance, to read a paper and then link directly to a related video?

AH: Exactly.

RP: By simply clicking on a hyperlink presumably?

AH: That's right. And since everything will reside in INSPIRE this will be a very simple process.

Central vs. distributed

RP: Ok, that's an important point I think. You are saying that the plan is not just to aggregate HEP content, but to host everything in INSPIRE. Rather than being sent off to external servers when they click on a link, therefore, users will invariably be directed to another file in INSPIRE?

AH: The future plan is to host everything that is freely accessible in INSPIRE, but our primary goal right now is to reproduce all the important features of SPIRES, not introduce masses of new content or features. It is very important that our users continue to feel at home and do not suffer disruptions. Nevertheless, even at this early stage they will notice improvements: increased speed, and easy handling of large sets of research results for instance.

The next step will be to make INSPIRE a repository for all Open Access publications. We envisage OA journal articles and OA conference proceedings will be harvested by INSPIRE, as CDS does now. Other material such as theses would also find a natural place.

RP: But for the moment the emphasis remains on metadata rather than full text right?

AH: Metadata is certainly the basis of what we are doing today. At the same time our user survey has shown that what scientists really want is access to the full text. The guiding principle will be to facilitate this, offering links to the pre-print and post-print OA material as well as the publisher version, as done by SPIRES today with the advantage that the OA material might be less clicks away.

RP: At this point, then, when users hit a link to a paper they may find themselves being sent to the publisher's web site, where I assume they are likely to be asked to pay to access a paper?

AH: If their library does not have a subscription. But they will also get the information whether a preprint is available — for free.

RP: Ok, to sum up: the content in INSPIRE will be founded on the bibliographic data currently hosted by SPIRES but the aim over the long-term is to create a full-text service — initially by adding full-text papers from CDS, but also encouraging OA publishers to archive the HEP papers they publish in INSPIRE?

AH: And actively harvesting all material we are entitled to. We also plan to add full-text mining capabilities, and the ability to find related articles.

RP: Will HEP scientists also be asked to deposit their preprints in INSPIRE?

AH: HEP scientists have already a full commitment to self-archive their preprints in arXiv and there is no reason to change this. What INSPIRE will offer is the possibility to deposit material that users want to be preserved in addition to today's possibilities, such as theses or old preprints which were not submitted to the arXiv originally: indeed, authors are usually happy to make them available but arXiv is seldom used as such a forum.

RP: The model you are adopting is an interesting one. However, a lot of people might argue that it is an outdated approach. The point about the Web, they might say, is that it calls for a distributed model. Indeed, that is the model inherent in the Open Archives Initiative (OAI-PMH), which assumes that records are physically distributed all around the Web not in a single repository. By using the OAI protocol repositories can then be aggregated into a single virtual archive by third-party services like OAIster. Harvester services like OAIster allow users to search the entire corpus of research papers through a single interface, regardless of where they are physically located. INSPIRE, by contrast, aims to host everything itself?

AH: And there are several good reasons for doing it this way. One is that it allows INSPIRE to also provide an archival service. After all, if you have lots of sources you have stability problems. Conference web sites, for instance, are notorious for their short lifespan. Most of the grey literature is OA but is in danger of getting lost if we don't archive it. At the same time, Invenio offers OAI-PMH and XML harvesting capabilities, so everything in INSPIRE will be harvested back by whoever might have an interest in it. Today that's the way a few IRs get their material: back from arXiv. We are lucky, in HEP, that authors recognise the importance of OA and their first port of call are naturally disciplinary and not institutional repositories. So, we're following a successful tradition!

RP: And like SPIRES you will curate the information hosted by INSPIRE too?

AH: Yes, we will continue to curate and enrich the data.

RP: It is very much a librarian's approach you are taking. The assumption is that curation and preservation are as important as providing access?

AH: That's true.

RP: The central vs. distributed debate is one that is hotly discussed within the OA movement. People frequently argue over whether it is better to adopt a distributed model in which you create thousands of institutional repositories, or rather build a smaller number of subject-specific central repositories like arXiv and — in the biomedical and life sciences arena — PubMed Central. With INSPIRE you are emulating the approach taken by arXiv. Others insist that a distributed approach is far better. Do you think that these two approaches will co-exist, or do you expect the centralised model to gradually elbow out the distributed model?

AH: Well, it's true that INSPIRE will be a central subject-specific repository, but I see no reason why institutional repositories cannot co-exist with a centralised model. As I mentioned, CDS will continue to operate as an institutional repository, and it makes sense that institutions will want to archive their own material. The important thing will be interoperability.

RP: And as you say, the Invenio software is OAI-PMH compliant.

AH: Correct.

RP: What will perhaps help is that the developers of the SWORD protocol expect to make it possible for repository managers to provide an automatic feed through of deposits made in institutional repositories into discipline repositories and vice versa. But tell me, does DESY have its own institutional repository?

AH: It does.

RP: Are DESY researchers mandated to self archive?

AH: We ask people to self-archive, and we send out e-mails to them reminding them to do so. We certainly want to fill our repository. And we have been quite successful at this. We have nearly 100% compliance in particle physics today. However, we have been less successful with our photon scientists.

RP: DESY researchers are requested to self-archive, but not mandated to do so?

AH: There is a mandate in the particle physics part of DESY. In fact, there was a mandate to self-archive in arXiv prior to our creating our own repository.

RP: When arXiv began it was only intended for a subset of the physics community. It then expanded out to cover all of physics, and today it includes some biology and maths too. Can you see INSPIRE broadening out in a similar way?

AH: We want to keep the advantage of being subject specific, but if we did expand into, say astrophysics, we would do so in a way that was driven by the community, and its interdisciplinarity. Anyway, for the moment, the plan with INSPIRE is to have two levels of records. There will be an identifiable inner core of material that is truly HEP and will be carefully curated

RP: Given the increasingly interdisciplinary nature of HEP then it will be important to define boundaries early on?

AH: It will. If you study the HEP publishing landscape you find a growing outer area of articles from adjacent fields that are of relevance to HEP — fields, like astrophysics, mathematics, condensed matter and so on. Identification of the HEP core articles is very important e.g. to ensure low noise in search results or enable statistical analysis of the HEP publication culture.

The outer area of HEP related articles will experience a lower level of curation within INSPIRE, but we plan in the long run to include all papers that have been cited at least once by a core paper. This is already done within SPIRES, but with a higher threshold — about 50 citations. With INSPIRE we expect over time to take that down to one citation.

Open Access

RP: We discussed the fact that the HEP community has in effect been practising Open Access for 40 years. But it might be useful for you to say how you define Open Access and what its objectives are.

AH: The definition of Open Access that I subscribe to is research that is not only free to read but free to re-use within the boundaries of fair conduct. That means that it is not only important that it is readable to the human eye, but to machines too. As to the objective of OA, I would say that it is to maximise the distribution and impact of research, especially research that has been publicly funded.

RP: And presumably it implies making research freely available on the Web from the moment of publication?

AH: That's my view yes. I'm also pretty confident that Open Access will be the standard of the future for scientific results.

RP: Would it be accurate to say that everything you are doing with INSPIRE is based on the assumption that in the future all research will be Open Access?

AH: Well, INSPIRE does not depend on Open Access; nor does SPIRES. But both databases are committed to an OA philosophy since they provide Open Access to metadata. And as we discussed, our dream is to get all the full text into INSPIRE too.

RP: What's for sure is that OA raises issues for publishers. But if HEP researchers developed a preprint culture very early, and have been practising OA for forty years, I am wondering what role publishers play for the HEP community in terms of scholarly communication. I was struck, for instance, that the report on the survey you mentioned earlier says that the main finding was that "community-based services are overwhelmingly dominant in the research workflow of HEP scholars". And citing the same report the press release announcing INSPIRE says that HEP scientists' information needs are "not met by existing commercial services". Presumably commercial publishers missed an opportunity here. Why?

AH: Well since particle physicists have always had a strong preprint culture, the main disadvantage of the commercial systems is that they don't cover preprints and they only cover a few conferences. That means that while our photon scientists are perfectly happy with commercial services like the Web of Science and ScienceDirect, our particle physicists have limited interest in them. The other point is that commercial services try to cater to the needs of all scholars, whereas the big advantage of community systems is that they can be tailored to the needs of their community.

What is quite unique about the particle physics community is that for many years it has organised its own subject specific database.

RP: SPIRES.

AH: Yes, and it is also significant that in doing so it has tried to be as openly accessible as current technology allows. So there has always been an OA philosophy amongst particle physicists, even before there was an OA movement in fact.

RP: So do physics publishers like the American Physical Society (APS), Elsevier, and the Institute of Physics (IOP) play any role in the HEP information landscape?

AH: Well apart from Elsevier's ScienceDirect they don't play any role in searching, and they are practically never used for that purpose. The main service they provide is in organising peer review; plus they have an archiving role.

RP: For particle physicists, then, publishers are viewed as the people you turn to when you want to have your research published, not when you want to read research published by your colleagues?

AH: For sure they do not look to publishers' sites or journals to search for information. Once they have found their information elsewhere, they click on the arXiv link if the article is unpublished (or they are at home). If it is published, it appears as a matter of taste: some would click on the preprint; some would follow the link to the journal.

I should add that the peer review service provided by publishers is much appreciated by the HEP community, and is widely considered an indispensable service. For that reason we cooperated extensively with publishers as we developed our plans for INSPIRE.

RP: In a paper you co-authored in May you described scholarly journals as the HEP community's "interface with officialdom". This goes to your point about providing credit: publishers offer a valuable third-party service to allow HEP scientists have their work validated and certified for external agencies such as funding organisations?

AH: Exactly. And having an independent authority validate your work is also very important when it comes to grants, evaluation, careers etc.

SCOAP³

RP: I believe you are on the SCOAP³ working party?

AH: I was. Up to the point where the working party delivered its final report in April 2007 and entered the fund-raising phase, I worked with Helmholtz, TIB and MPG to organise the German participation to the consortium.

RP: Can you say something about SCOAP³, and what you hope to achieve with it?

AH: SCOAP³ is a consortium of HEP institutions, libraries, and funding agencies; and our aim is to convert the entire HEP peer-reviewed literature to Open Access — in collaboration with publishers of course.

RP: How do you plan to achieve that?

AH: We propose a business model in which the current journal subscription funds paid by research institutions are redirected, via a central fund managed by the SCOAP³ consortium, to pay for the peer-review services that publishers provide. The aim is to make all HEP journal articles free to read and reuse as the community wants. In the process, we hope to alleviate the serials crisis.

RP: So the HEP community is seeking to pool all the journal subscriptions it currently pays to publishers and, by acting in concert, plans to change the nature of its relationship with publishers, and ensure that all future research is made OA as a result. And in the process of removing access charges you hope to reduce the overall costs of scholarly communication?

AH: Yes.

RP: How much money do you think can be saved?

AH: That is very hard to say, because at the moment the market lacks transparency, and nobody really knows what the community as a whole pays for a journal. It is hidden in all these big package deals.

With SCOAP^3, however, publishers will be asked to tender for the service they provide and in doing so they will have to specify what compensation they expect for that service. This will mean that for the first time we will know exactly what the community pays for publishing a paper.

RP: Clearly what it pays for a paper will depend on the outcome of the tender. In an article CERN's Scientific Information Officer Jens Vigen wrote for Research Information last year, he indicated that you anticipate a figure of between 1,000 and 2,000 Euros per paper.

AH: Jens refers to the final report of the SCOAP3 working. There won't be just one price for every journal article: it may be related to criteria like the quality and reputation of the journal or its rejection rate.

RP: Some would argue that 1,000 to 2,000 Euros is still far too high and that it could be much lower. You don't agree?

AH: This price was estimated from publicly available numbers, such as those from the APS (the American Physical Society) which publishes about half of the articles in the field.

RP: As I understand it the central fund you aim to create is expected to be in the region of 10 million Euros. How will you calculate what any specific research institution should contribute to that fund?

AH: The financial burden will be distributed evenly among all countries using a fair-share model based on the current distribution of articles per country. SPIRES today and INSPIRE tomorrow will provide the statistical information we need to calculate each country's contribution.

RP: Ok, so the more HEP papers an institution currently publishes the greater will be their contribution, and vice versa. Can I check my understanding: there are currently 10 HEP journals, 5,000 HEP papers published each year, and 20,000 HEP scientists — who are funded by 50 funding bodies. Is that accurate?

AH: That's roughly correct, but on a country-by-country basis, not necessarily at the institutional level. HEP articles are published in more than 100 journals. But our task is made easier given that just 6 or so journals carrying almost exclusively HEP content cover about 80% of the literature. As for the number of scientists it depends where one draws the line of interdisciplinarity, it could be up to 40,000.

RP: But those are the figures you are working on?

AH: Yes, although the figure of 5,000 journal articles doesn't include conference proceedings that are published in journals and which are currently not in the focus of SCOAP^3.

RP: So in effect, the SCOAP^3 consortium plans to sign up the whole HEP community and then go to the publishers and say, "We want to re-negotiate our relationship with you, and to that end we invite you to bid to do what you are already doing but on different terms"? Presumably you don't expect to commission one publisher to do everything: you would expect to commission a number of publishers?

AH: We hope to get all the important HEP journals on board, so yes there will be several publishers involved. We hope that all of them will answer the tender, and subsequently agree to a contract based on conditions that are agreeable to both sides.

RP: You are assuming then that the same publishers will continue publishing the same journals, but on re-negotiated terms and conditions, and at a price, that better meet the needs of the HEP community?

AH: One could put it that way. The real novelty is that quality, price and volume will be related. And the tender won't be restricted to well-established journals. We will be open to new journals as well. And perhaps SCOAP^3 will alleviate the discrimination small publishers currently suffer from the big deal model.

RP: In what way does the big deal discriminate against small publishers?

AH: It means that in times of tight budgets librarians find it easier to cancel journals from smaller publishers.

RP: You mean that it is much easier for libraries to drop a small publisher who offers just one or a small number of journals than it is to re-negotiate the contract with a large publisher that might encompass hundreds or even thousands of journals?

AH: Exactly. Often contracts with large publishers are long-term contracts with clauses restricting subscription cancellations. So in my view SCOAP^3 will offer more of a level playing field.

RP: In answering the tender then a journal will go to SCOAP^3 and say, "We'd like to offer such-and-such a journal, and this is what we propose that SCOAP^3 pays us each year for peer reviewing all the HEP papers we publish, and making them freely available on the Internet?"

AH: Correct. And then we will compare the bids we get from different publishers.

RP: Might some publishers be turned away?

AH: We are hoping that all publishers of high-quality journals carrying HEP content will be selected. The limiting factor is the total budget envelope of 10 Million Euros and therefore the price per article which will be quoted in each bid.

RP: You mentioned the serials crisis earlier. This refers to the fact that subscription-journal prices have risen to the point where even the wealthiest universities in North America can no longer afford to buy access to all the journals their faculty need. I wonder if the danger with the SCOAP^3 approach is that current prices will simply be locked into a new OA model. Even if you tender for the services publishers provide, how certain can you be that that publishers will propose a fair price? Indeed, it's possible that they might not even know themselves how much it costs them to manage the peer review process, and they certainly won't want to quote a price that would produce less revenue. How can you guarantee that once the process is over the HEP community will be paying less than it currently pays for having its papers published?

AH: SCOAP^3 will inject competition and transparency in the market, linking quality and value. This will for sure have economic benefits. However, our primary goal is Open Access publishing, and the integrity of the peer review service. SCOAP^3 tries to establish a sustainable OA business model in the least disruptive way for authors, readers and publishers.

RP: Achieving Open Access is sufficient in itself then?

AH: It would be a tremendous step forward to convert the journal literature of a whole scientific discipline to OA.

RP: You said that the main criterion for choosing which publishers to commission will be the quality of the journals they offer. You also said that you are open to bids from new journals. Presumably one problem that new journals will face is that they will not have a track record, making it hard to assess their quality?

AH: We are currently working on several possibilities to assess the "quality" of a journal which should account for this. One criterion for a new journal could e.g. be the reputation of its editorial board.

RP: I'm assuming that SCOAP^3 will only work if you can get the entire HEP community to buy into it. Is that correct?

AH: Indeed, and not only the HEP community but the bodies which today purchase subscription to HEP journals. In some cases these are HEP laboratories or large funding agencies, in most cases these are libraries and library consortia. Our success sits with unanimity, not majority. You can say that's a weakness, but we think that's our strength. After all, we have already collected almost half of the funds in one year.

RP: Are you confident that you will reach critical mass on this?

AH: The signs are positive. After all, in one year with a model which goes against three centuries of tradition, we've already enlisted half of the partners

RP: If you do, when do you expect the initiative will be launched

AH: It would be great to have it in place for the first publications describing the discoveries at the Large Hadron Collider (LHC). After all, the LHC experiments have voted statements committing them to publish Open Access, and in one case mentioned that they will "privilege SCOAP^3 friendly journals" for their publications.

RP: Presumably you are already talking to publishers?

AH: Oh, yes — we have been talking to them from the beginning. They were already involved in the Task Force of Open Access Publishing in Particle Physics that proposed the development of the SCOAP^3 model in 2006.

RP: But there will be an official tender process at a later date?

AH: Correct.

RP: Who will have responsibility for the preservation and archiving of HEP literature if SCOAP^3 does succeed?

AH: Most of the system will look like today. So publishers will continue doing their part, but thanks to OA, repositories like INSPIRE will also get the content

RP: Some have expressed concern that one outcome of SCOAP^3 could be that the whole physics literature will be forced into an OA author-pays model, and those researchers who are not part of a large research institution like CERN or DESY will find themselves having to pay to publish their papers out of their research funds. Is this a genuine concern?

AH: No, this is not correct. The idea of SCOAP^3 is that authors will NOT pay directly for their publications, whatever their institution. All authors will be able to publish wherever they see fit, as today. The peer-review process will be just paid in a different way, and everyone will have access to the final peer-reviewed articles. SCOAP^3 is a community-pays model, not an author-pays model.

RP: So you don't believe that there is any danger that some researchers will have to pay for the publication of their papers from their own funds?

AH: No, not at all — since all publication charges for articles appearing in journals collaborating with SCOAP^3 will be centrally paid by the consortium. Moreover, we plan to set aside a portion of the SCOAP^3 budget expressly for the purpose of covering the costs for low-income countries who cannot be expected to contribute to SCOAP^3 at this stage.

RP: Do you think the consortium approach utilised by SCOAP^3 is only relevant to the HEP community, or might it be a useful model for the wider research community? Can you see it being successfully carried over, for instance, to domains like condensed matter physics, fluid dynamics, theoretical astrophysics, plasma physics, or to oceanography, or plant biology?

AH: I definitely see the possibility to carry it over to other disciplines with similar characteristics: Small, tightly-knit communities where authors and readers are largely identical. I can easily see SCOAP^3 extending to nuclear physics and astrophysics where there's already a large overlap with particle physics.

RP: There are some who believe that the SCOAP^3 model cannot be generalised to other domains. Consequently, they argue, there is a danger that other disciplines may try to emulate SCOAP^3, only to get into difficulties. If this were to happen, they add, it could have a negative impact on the development of OA. Do you think this is a genuine concern?

AH: We hope to give an impulse to other disciplines by demonstrating the viability of a new OA business model. But every interested community has to — and certainly will — investigate whether our model can be adapted to its specific needs.

RP: One university-based physicist also suggested to me that while the SCOAP^3 model may work well for large HEP research institutions like CERN and DESY, it is unlikely to work in a university environment, where HEP is only one of a number of disciplines, many of which are funded in a different way to HEP. Does he have a point? Is it possible for SCOAP^3 university partners to operate in one mode with respect to particle physics, but in quite another with respect to other disciplines?

AH: As I said before, in some countries the partners of SCOAP^3 are HEP research institutions or funding agencies, but the vast majority of our partners are indeed universities, through their natural federation in consortia and the natural process through which they purchase journals. It is a matter of re-directing the subscription funds.

RP: As I understand it, HEP is an almost 100% arXiv discipline. In other words, the norm within the HEP community is for researchers to publish in a subscription journal and then self-archive their papers in arXiv, thereby making them freely available to their colleagues without the need to publish in an OA journal — a practice referred to as Green OA. The same point was made to me by CERN Director General Elect Rolf-Dieter Heuer when I interviewed him for* Computer Weekly *recently. As he put it: "particle physicists have had OA for many years now". If that's correct, then why does the HEP community need Gold OA, and why does it need SCOAP^3? It already has Green OA. Isn't that enough?

AH: The normal way is the other way round. Most authors send their papers to arXiv before or parallel to sending it to a publisher. Green OA is a great achievement. But the problem with it is: If you look at a paper in an institutional or subject repository you seldom can be sure that you are really looking at the final, published version. Since a paper may undergo quite substantial revisions in line with a referee's suggestions it is essential to have access to the peer-reviewed version. With SCOAP^3 we also want to put the peer-review system on a solid basis. The community is worried that with increasing prices and increasing cancellations the journals they need could be threatened.

Open Data

RP: I believe you also take an interest in Open Data. Will that be reflected in the development of INSPIRE?

AH: Well, one thing we are planning to do with the papers deposited in INSPIRE is to make the data behind the figures and tables in them usable. Right now scientists sometimes still have to sit down with a printed copy of a paper and use a ruler to take data points from the graphs published in them.

RP: So you would mine the papers deposited in INSPIRE, harvest the data in the published tables and graphs — presumably by means of screen scraping — and then make that data available in computer-readable form? This is the kind of thing that Peter Murray-Rust is doing with journals in the chemical field isn't it?

AH: That is one possibility. However, we think a more elegant solution would be for the scientists themselves to provide the data in a suitable form.

RP: So when they self-archived a paper in INSPIRE you would encourage researchers to also deposit the tables and other numerical information in a machine-readable form. This would allow you to link directly from the paper to the data?

AH: Papers will not be self archived in INSPIRE, as I said before, but we could plan to work towards a system to allow an easy exchange of deposited material between authors and systems and between systems.

RP: In defining Open Access you also stressed the importance of reuse, and the need for research to be machine-readable. Clearly, then, licensing is important. This is perhaps one good reason for preferring Gold OA over Green OA: Most OA publishers use Creative Commons licences that permit re-use; subscription publishers generally do not, and so self-archived papers published subscription journals may not be reusable.

AH: That's right.

RP: This is an area where Murray-Rust has had considerable problems — because even though facts cannot be copyrighted, subscription publishers tend to assume that they acquire ownership of everything they publish, and some have therefore threatened to sue researchers who extract data from papers. It is for this reason that Murray-Rust has become such a passionate advocate for the so-called Open Data movement, which like you maintains that scientific papers need not only to be freely available for humans to read, but for machines to read too.

AH: And, as you say, if the papers are OA the problem goes away.

RP: Of course legal barriers are just part of the challenge. Extracting data from a PDF file also raises technical issues. For this reason Murray-Rust intensely dislikes the PDF format, and helped develop an XML-based system for chemical information called Chemical Markup Language (CML) to avoid such problems.

AH: I completely agree with Peter that PDF is an outdated format. We have to come up with something better for the future.

RP: Do physicists yet have anything similar to CML?

AH: Not at the moment, but it is certainly something we are interested in developing, in collaboration with publishers.

RP: It does not help that PDF has become the standard format for publishing scholarly papers I guess, and most papers deposited in arXiv today are in PDF. However, I notice that there is

often an associated PostScript (PS) file as well, which I think is not typical with self-archiving. Where would PS fit into a post-PDF XML-based environment? Would there still be a need for PS in fact?

AH: The PS format is a page description language developed for printing and for many printers PS is the "native" language. This means that even when you print a PDF file it is most likely that the PDF is first transformed into PS and then printed. For that reason, PS will play an important role for printing self-archived physics papers in the foreseeable future.

RP: Can you say more about the difference between PS and PDF?

AH: PS is just a page description language for printing documents. PDF, by contrast, was designed not only for printing, but also for viewing. Since PDF is also a page description language, the visual output of a PDF file is independent of the device being used for displaying it, so it can be viewed as an electronic version of paper. (And with the 2005 PDF/A ISO standard it is also now suitable for long-term archiving purposes.)

So while PDF is fine for human readers, all layout-oriented formats share with paper the disadvantage that structural information — if contained at all — is at most an add-on, but not inherent to the format. So, for example, the notion of a chapter or a section might or might not be present, whereas e.g. the relation between symbols representing a formula is always lost. This makes it very difficult if not impossible to reconstruct the full structural information automatically so that machines can use it.

RP: And it is because machines cannot do much with a PDF file that Murray-Rust so dislikes PDF. An added complication for physicists, perhaps, is that in order to express mathematics in their papers they have long used the TeX typesetting system. Would it be easy to incorporate TeX into an XML-based format?

AH: TeX and especially LaTeX (which is a TeX Macro package) are focused on the structure of documents. This means that TeX/LaTeX source files contain virtually no layout information. The layout information is contained in additional files, including the TeX program. This makes it easy to transform them into various output formats like PS, DVI (an intermediate binary output format of TeX including layout information), and PDF etc.

RP: Just to clarify things: the distinctions you are drawing here are between layout information (which PDF majors in), page description information (which both PS and PDF provide), and structural information (which Tex and LaTeX offer).

AH: Right, and since TeX is a fully-fledged programming language it always needs the corresponding compiler (e.g. the TeX program) to produce visual output. This has implications for archiving. And to make things even more complex there are a huge number of additional "standard" packages in use, which are needed to successfully "compile" a TeX/LaTeX document.

RP: If one wants HEP papers to be machine readable then perhaps the key question here is whether it is possible to produce an XML-based format that can handle Tex and LaTex?

AH: Well, to date all programs that have been written to transform TeX into XML have flaws that prevent a 100% success rate. The hope is that TeX/LaTeX can be transformed into an XML/MATHML/XHTML format. The advantage of the latter is that it is structured, human as well as machine readable and it allows additional information — such as data, program code, multimedia information etc. — to be incorporated into it. This isn't possible with PS, PDF/A or TeX/LaTeX.

So while data and program code can easily be formulated in XML, for instance, this might be quite difficult or even impossible for other additional information. That means that these must either be included in their "native" form — with all disadvantages for long-term preservation — or a still mightier format must be created.

RP: To sum up: the issue of Open Data encompasses both legal issues (which would be solved if Open Access publishing became the norm) and technological issues — essentially how do you

make data machine-readable. When I interviewed Rolf-Dieter Heuer for* Computer Weekly*, however, he seemed to be concerned about Open Data for a quite different reason. In talking about the LHC, for instance, he pointed out that the collider will produce huge volumes of data (15 petabytes a year). Given that the LHC has cost 6 billion Euros in public money, he suggested, it is important that this data be reusable. This is not a concern about data published in papers, but the raw data that particle accelerators produce.

AH: Well, INSPIRE is independent of the LHC and I am not involved with that project. However, it is clear that the HEP community needs to find a way to make experimental data reusable. The problem is that while there is no reason not to make HEP data Open Access — in fact we might feel a moral obligation to do so — at the moment we lack the ability to do so.

RP: Why is that?

AH: It's not so much an issue of the quantity of data produced — after all astrophysicists deal with similar volumes and have a tradition of making their data public — but the sheer complexity of HEP data. There is also the issue of the amount of effort that would need to be invested to make the data reusable.

RP: When you talk about the complexity of HEP data you are referring to the fact that it is not enough simply to provide access to HEP experimental data; you also need to provide a lot of background information — precise descriptions, for instance, of the conditions under which the data was collected; because without this background information it is not possible to interpret the data in any meaningful way?

AH: Correct. So we will have to develop what we call a parallel format — that is, a format that not only preserves the data itself, but also the necessary knowledge to be able to interpret it.

RP: Essentially you are saying that when you preserve the data it will be necessary to embed knowledge about it (How it was captured, on what basis it was calculated, how to interpret it, etc.) into the data itself?

AH: Correct. At the moment this knowledge is hidden in the heads of people and gets lost when an experiment is disbanded. So we are going to have to create some high-level objects to capture this information and preserve it with the data.

RP: Superficially it sounds like a simple metadata issue. However, I suspect it is more complicated than that.

AH: It is, yes; because we are talking about all the knowledge about the data in people's heads.

RP: It's essentially a knowledge management issue then?

AH: Not exclusively. It's also a financial issue — because while it would cost only a minute fraction of the capital investment of a big experiment to fund it, a small fraction is still a big number in a 6-billion-euro project like the LHC.

In addition, there is a sociological barrier to data preservation. Spending time on it, for instance, competes with actually doing research. And today there is very little academic incentive to devote time to such matters. So it is not primarily a technical problem. After all, data migration is quite common.

RP: But from what you say, ensuring that the data is regularly migrated to new formats as the old ones become obsolescent is only the first step.

AH: What happens is that our computing centres migrate data from one format to another, but the information to understand the data can be easily lost forever just after the data is taken.

RP: To go back to the high-level objects you mentioned. What would they consist of, and what sort of things would be contained in them?

AH: They would contain all the knowledge you need to interpret the data. And, as you say, this information will need to be packaged with the data itself. But this is a task for the experimental physicists, working with the IT people, not for library-based information professionals like me. And right now they are just at the beginning of the process. The people conducting the experiments are going to have to sit down with the IT people and work out how to do it.

RP: So for the moment they are still at the point of deciding what information the objects will need to contain, rather than working out how it can be done technically?

AH: That's true, but you know this is a problem that extends well beyond HEP, or the physics community at large. To that end CERN is currently involved in an EU FP7 project called PARSE.Insight.

RP: What is PARSE.Insight?

AH: It's a project that has just started and which is focused on understanding the future of preservation and the reuse of data. It is closely associated with an organisation called Alliance for Permanent Access to the Records of Science, which is looking at ways to create an infrastructure for permanent access to scientific data and publications.

RP: Can you see INSPIRE playing a role in any of this?

AH: Well, the LHC will produce a huge number of publications, and we need a modern platform that can deal effectively with the flood of information we can expect — which is one of the main reasons for developing INSPIRE. In addition, you could argue that INSPIRE would be the natural place to look for the high-level objects we discussed.

But as I say, this is mainly an issue for the experimental physicists right now, and for those responsible for the LHC Grid, by means of which the data will be distributed and managed.

Futures

RP: Nevertheless, perhaps one could foresee a future role for INSPIRE in the larger revolution that encompasses things like eScience, Grid computing, the Cyberinfrastructure the Semantic Web etc. etc. One could certainly envisage that the experimental data generated by "Big Science" will need to be linked to the research papers produced by analysing that data somehow. And I think we have agreed that both types of information will need to be interoperable, machine-readable and reusable.

AH: As I say, I am not personally involved with any of these other developments for the moment. INSPIRE and the LHC Grid are not connected in any way currently. But there is a potential to connect repositories and the Grid. Many great things repositories could do with metrics, bibliometrics, text mining and data mining are computationally intensive. And if they were made interoperable with the Grid they would be relieved from the shoulders of repository managers who could offer more services without technical bottlenecks. In so far as it is a central component of knowledge management, INSPIRE is part of the e-Infrastructure. But it is too early to say how INSPIRE will be integrated with other components.

RP: But you could envisage a role for INSPIRE in this?

AH: Sure, we could soon be thinking of something along these lines.

RP: The rise of large central repositories like INSPIRE, arXiv and PubMed Central, plus the constant growth in institutional repositories, suggests to me what we are rapidly moving away from a journal-based model of scholarly communication towards a database model. Would you agree?

AH: I agree absolutely.

RP: What are the implications of that?

AH: The main implication is that people will find what they want more easily. This development is part of what some people call the deconstruction of the notion of a document, and it means that the classical journal article will not remain the main vehicle for scholarly communication in the future. It means that we will see different materials and different media developing corresponding to the different stages of the research process. The key point is that these developments just don't fit into the journal model.

RP: Which brings me back to the question of where INSPIRE fits into the big picture. If scholarly communication is moving to a database model it opens up the possibility of institutional or subject-based repositories becoming the primary publishing platform — which some predict will happen. In this model publishers become service providers who manage the peer review process on an outsourced basis. You could argue that SCOAP^3 represents the first step in that process. It is also the model that the University of California is promoting. Do you see scholarly communication developing in that way?

AH: Underlying these models is the postulate that scientists should retain ownership of their work — to which I fully subscribe. I see SCOAP^3 as a catalyser in this context, by answering two basic needs of the HEP community: access and high-quality peer review.

RP: Perhaps the next step will be the development of overlay journals, where researchers self-archive their preprints in repositories and publishers are then invited to peer review the papers — by, for instance, developing virtual journals that simply link to the peer-reviewed versions of the papers, which are hosted in different repositories.

AH: It is not currently on the INSPIRE agenda to become a publishing platform. However, I can see that INSPIRE would make an ideal test bed for such experimentation. Our aim, after all, is for INSPIRE to eventually host the entire corpus of HEP research.

There was, by the way, an early example of an overlay journal in our field; *Advances in Theoretical and Mathematical Physics*. It hasn't been overly successful, but that may simply be because it was too early in the game.

But we shall have to wait and see. For the moment our mantra is: "Build it, put it online and develop it in line with the wishes and needs of the community."

RP: But you are sympathetic to the University of California's vision? Certainly the UoC talks about taking back ownership of research, and it seems keen to reinvent publishers as providers of outsourced peer-review services too.

AH: Well, for the moment I am quite happy for arXiv to disseminate papers, INSPIRE to be the search interface and the journals to do peer review in the way they have traditionally done, so long as they make their papers OA — and for the moment SCOAP^3 is focused on supporting that model.

RP: It occurs to me that instead of developing INSPIRE you could have gone to Paul Ginsparg and said, "Why don't you develop the level of search functionality available in SPIRES, and let us pension the service off?" Doing so would have provided the HEP community with a single resource to meet all their needs, rather than two. Perhaps this goes to the issue of ownership: large research institutions like DESY, SLAC and CERN want to ensure that they retain ownership of HEP research, and they feel that the only way to do that is to develop and own their own repository service?

AH: It is not mainly a question of ownership. We want to cater to the specific needs of our community. The thematic scope of arXiv is much broader than HEP. At the same time it has a much narrower scope than INSPIRE since it is a preprint repository. And we want to cover all relevant

material, going much farther back in time than 1991, curating and enriching the metadata. But I'm sure we will witness a growing symbiosis between INSPIRE and arXiv — and possibly ADS, the central publication database of the astrophysicists.

RP: Can we peer into the future a little more. How do you expect physicists to be sharing their research with one another in, say, ten years, and where do you expect publishers to sit in that process?

AH: I think HEP will pretty much continue on the path it is taking today: it will remain a preprint culture, and it will probably put more emphasis on other forms of scholarly communication, including things like conference notes, slides and grey literature. And while I think the classical peer-reviewed paper will survive, it won't be the dominant form of scientific communication any more.

RP: You talk of new kinds of media developing. Can you say more about this?

AH: Well, one point to make is that this is not just about making data reusable: it is important that all kinds of arcane knowledge become visible too — the open lab notebook is one possible channel for doing this.

RP: What do you see driving this development?

AH: It is important to make all this hidden knowledge visible since it could offer useful tricks, or help scientists to understand results better or to avoid repeating mistakes.

Another development I find fascinating is literature-based discovery, and whether and how it can play a role in particle physics. Do you know about the way in which the biochemical pathway in drug addiction was discovered from the literature?

RP: You are referring to the way in which a group of Chinese researchers made a scientific discovery without doing any original research themselves, but simply mined published papers?

AH: Correct. They made their discovery extracting data from more than 1,000 scholarly papers. I'm curious whether comparable discoveries might be possible in HEP. This is an area of interest to an international collaboration in particle physics called the particle data group.

RP: Can you give me some examples of how in the future HEP researchers are likely to a) publish their research and b) locate and read research produced by their colleagues?

AH: I imagine a colourful landscape of different publication models — all of which will be Open Access. I envisage, for instance, virtual journals overlaid on arXiv or INSPIRE.

I also expect publication to become a much more collaborative effort, following the examples of chemistry and economics. And I can see journals practising open peer review, inviting readers to comment on a paper at a very early stage in order to improve the final article, for instance. Publication would then consist of combining the paper with the comments that were received and the referees' reports.

Even more radically I can see a piece of research starting by a scientist simply putting an idea on a wiki, where it would be time-stamped in order to establish precedence. Other researchers would then come along and elaborate on the idea, or write a program or do some calculations to test it, or maybe visualise it in some way.

In this case publication would consist of pulling all the pieces together — all of which would be independently citable. And this process of aggregation could be done by a journal. Then, even after publication, the aggregated "paper" could be further developed by researchers adding further comments and links to related material etc.

RP: And if, as you suggest, all this research was Open Access I guess we could expect to see the development of new tools that were able to aggregate it automatically, and then perhaps use it to create new value — using similar techniques to literature-based discovery for instance?

AH: Absolutely. In addition, we will need to design new quality measures. We have relied much too heavily on the impact factor of journals, so perhaps we can also find a better way of measuring the quality of the new materials that become accessible. In fact, this is another reason for making more hidden information visible.

RP: How do you mean?

AH: The current credit system is inefficient and already out of date. We need, for instance, to redefine the definition of authorship. After all, when we have a big experiment nowadays the papers that are later published may have thousands of authors credited.

RP: Yes, I'm told the record to date is 1,113 authors. Your point is that when you have so many authors contribute to a paper it is very hard to apportion credit properly or fairly?

AH: Actually there are 3,078 on the (Open Access!) paper describing the construction of the CMS experiment. We have no way of telling what any particular author has contributed to the research. This means that when he or she applies for a job their list of publications may not be particularly helpful. In such circumstances the potential employers would be better off if they had access to the actual scientific work of the applicant.

RP: So they should be looking to things like wikis for this should they?

AH: To wikis, and to raw data and to software programs, or videos and blogs etc. So as more and more of this kind of material goes online it could be used for credit purposes. This suggests that some aspects of scholarly communication will no longer be characterised by a paper trail, but by an e-mail trail, along with wiki pages, blogs, virtual workbenches etc. What's important is to make it citable, and assessable.

RP: The challenge will lie in developing effective ways of aggregating the diverse material you refer to in order to arrive at a certification mark like the traditional impact factor or citation count?

AH: That is the challenge, yes — to aggregate it and to measure its impact.

RP: Earlier we mentioned the serials crisis, and you pointed out that no one really knows the true costs of scholarly communication, although you hope to establish this better with SCOAP^3. If we add to this the likelihood that the research community will need to devise new ways of assessing the work of a researcher then I wonder if perhaps we can we expect traditional peer review to eventually disappear? After all, there is a view that in an online environment the only significant remaining costs of scholarly communication are the costs associated with organising peer review. If, despite initiatives like SCOAP^3, publishing costs remain insupportable, would it not make sense to reinvent the whole process and do away with peer review?

AH: I think peer review will be needed for a long time to come, at least in our community. In fact, it could even be that the price of peer review will go up.

RP: Why?

AH: The HEP community wants to preserve peer review, so it may be that what we can save on other publishing costs might be redirected to improving the peer review process. SCOAP^3 is putting access, peer review and sustainability in the same equation. However, I think that peer review will diversify, and I hope we will see experiments with things like open peer review in HEP. If open peer review becomes the norm, by the way, I doubt referees will stay anonymous.

RP: Why?

AH: Because the HEP community is a very close community, which means that if a referee's report becomes visible his/her identity could probably be easily guessed by people working in the same field. So we would have to go all the way.

Chicken and egg

RP: As we agreed, particle physicists have in effect been practising Open Access for over 40 years now. Indeed physicists are widely credited with having invented Open Access. What is it that is special about the HEP community that caused it to develop a strong preprint culture, and at such an early stage?

AH: Probably because it is a small closely-knit community and the authors and readers are practically identical. The community has also always been part of a very international enterprise, and worldwide collaboration has long been the norm. This has meant that the rapid long-distance exchange of information has always been crucial — something that is not as important in many other disciplines, which becomes obvious when you compare citation histories of typical papers in different disciplines. In addition, of course, the HEP community has always been at the cutting edge of technology.

RP: So new ideas and discoveries occur frequently, which makes researchers impatient to know the latest developments, which means having access to HEP papers before they have been published in a journal?

AH: Sure, but remember that HEP researchers are not the only people to develop a preprint culture. Some other disciplines, like economics, have too. We should also note that a preprint culture does not exist in all areas of physics; it is mainly particle physics.

Even at DESY one can observe a cultural clash between the particle physicists and the photon scientists, who still rely completely on journals. And even within the particle physics community we find that the accelerator physicists are much more dependent on conference proceedings than preprints. In other words we are not talking about a homogeneous situation.

Nevertheless, I think it does say something about the mentality of physicists: If they encounter a problem they immediately want a solution. If nothing ready-made is available — a very common situation — their strong self-confidence and pronounced playfulness lead them to sit down and try to work it out themselves, and often with success.

RP: How much do you think the preprint culture has been driven by the technology, and how much has it been driven by HEP researchers developing new research tools like arXiv and SPIRES to satisfy their desire to exchange their research with one another as quickly and effectively as possible?

AH: That is a classic chicken and egg question! I would say that both things have gone hand in hand.

RP: But as we discussed, physicists have been ahead of the curve?

AH: Well, as I say, HEP scientists work at the cutting edge of technology, so it is very natural for them to find technological solutions to their communication needs.

RP: And as the LHC demonstrates, particle physicists also tend to be very technology-literate don't they?

AH: That's true, and so they were very early adopters of computer technology. In addition, of course, HEP has always been big science, we have these huge research centres like CERN and DESY. So if a physicist comes up with a creative new idea there is a good chance that he will find enough supporters within the research centre to realise his idea.

RP: And there will likely be enough money available to fund its development I guess!

AH: Yes, as you say, there may be a budget; and there will be a lot of local expertise on hand too.

RP: Ok, let's leave it there. Thank you very much for your time.
