Interview with a consultant in the field of long-term documentation at the *Bundesamt für die Sicherheit der nuklearen Entsorgung* (BASE) (September 25, 2024)
Length: 2:00h

PM:

Can you describe your activities and responsibilities at BASE?

Expert:

I am a consultant in the field of "long-term documentation," which is part of the research department. And I am the consultant for digital long-term archiving there. I usually use the term "long-term archiving of digital data / long-term preservation." We'll come back to the definition of terms in a moment. My background is this: I studied philosophy and computer science a long time ago and started working in the field of long-term archiving almost 20 years ago. First as a software developer and then in a series of research and development projects. This was mostly in the field of research and memory organizations, cultural institutions, and scientific institutions. On the one hand, in the library context. And for a while, I was also more involved with the topic of research data management. But always at the interface between IT, memory organization—which for me is the umbrella term for libraries, archives, museums, and so on—and science.

My job: The field is called long-term documentation. This generally refers to the preservation of information and knowledge in the area of interim storage and final repository. And long-term archiving is, to a certain extent, a German artificial concept. It's a discussion that has arisen particularly in libraries about how to keep digital objects usable. How to preserve them over a long period of time. And then the term long-term archiving was chosen without—to put it simply—without asking archivists.

Because if you ask an archivist, they will say: "There is no such thing as long-term archiving, because all archiving is permanent. There is no end date." This is important to archivists. That is why they usually do not use the term long-term archiving. To differentiate: if you come from an IT background, archiving is often understood to mean something like storage. For people with an IT background, the addition of "long-term" makes it clear that this refers to a period of time in which technological changes, among other things, play a role. And that is how it is usually understood in German. Long-term archiving is the preservation of the usability of digital information or digital data over a period of time in which contextual changes, technological changes, and social changes become significant. So long-term archiving always means digital. There is no such thing as analog long-term archiving in this sense. In this sense, it is a German technical term. And so you can also distinguish it a little from long-term documentation. For us, long-term documentation is simply digital and analog. And to be honest, I only know this term in the field of nuclear waste disposal.

In other contexts—libraries, archives, museums—I have never come across the term long-term documentation before. This means that long-term documentation is already a term specific to our field of work. And if we want to make one last pedantic clarification or refinement, then we can of course say that documentation refers more strongly to the process of recording something as it is being created – in other words, documenting it. Whereas archiving refers to documents that were created in the past and are then taken over and stored permanently. That is, so to speak, the conceptual field in which we operate.

PM:

Are there any English synonyms for this?

Expert:

In English, the typical term for long-term archiving is "long-term preservation." And long-term documentation... I'm not exactly sure what term is used in the international context. But I could look it up, because there are at least a few international committees or institutions that deal with these issues. But I would just look up what they use for it myself. But when it comes to digital information, the term in German is "Langzeitarchivierung" and in English "long-term preservation." Sometimes you have to be careful that when you use the term "digital long-term archiving," people understand it to mean digitizing documents and then preserving them. That's not the point of this term. But that's not the area of responsibility I'm concerned with. And I would say that this is a special case. To express it as precisely and free of misunderstanding as possible, I often say "long-term archiving of digital data / long-term preservation" or "long-term archiving of digital objects." Sometimes museums have certain reservations because they understand an object to mean something very specific. When it comes to the term "data," some people don't understand it in such a general sense and therefore don't include documents under it. But I mean it in the broadest possible sense.

PM:

What hardware and software do you work with in your environment? Do you have your own servers and how do you work with media?

Expert:

We are currently still in the process of setting things up. That's why there is little that is definitive – both in terms of the data to be stored and the infrastructure or tools we use. And that's also because there is this paragraph from the site selection procedure, § 38. And the second paragraph essentially says: "Everything else is regulated by a regulation." We are still waiting for this regulation and are, of course, preparing for it. We are currently preparing a tender for a long-term archiving system. That is why I cannot tell you yet which tools, products, or providers we will be using as instruments or infrastructure. This is because it is still under development and the tender is being prepared. The other thing is that what we usually call data, because that is also the term used in the law – i.e., for the sub-task of long-term archiving of digital data, the digital data that we will have to store – we will only receive that later. But we expect that we will be confronted with pretty much everything imaginable. The key points that were developed for the BMUV [Federal Ministry for the Environment, Nature Conservation, Nuclear Safety, and Consumer Protection] regulation on public participation included a list of significant data. It is not exhaustive, but if you look at it, it is more technical and content-driven. But of course, you can already deduce a lot from it about what this will mean. Of course, in terms of media type, it will be normal office applications, Office documents. But it also includes things like geodata. In other words, things that express geological facts. Building plans and similar items are also mentioned . And from this, we can already deduce that we will be confronted with formats such as those used in BIM [Building Information Modeling]. That we will be confronted with data that is generated during geological investigations of the subsoil, e.g., from boreholes. And that this will of course not only be in the form of individual files, but also in database form.

PM:

So that means not only archiving individual files, but also self-contained databases?

Expert:

That's a good question. I suspect that sooner or later it will come down to that. But as things stand at the moment, we don't know yet. In addition, of course, when the task begins in a few years, it will naturally continue for many years and decades. This means that even if we knew what the current status was, there is technological progress and we will be confronted with new forms that simply do not exist yet. That is part of the nature of this task.

PM:

Database management systems [DBMS] have also kept up with the times, I think. This raises the question: even if you have a well-structured database like an Oracle database now, will I still be able to read it in 50 years?

Expert:

Well, database preservation is an international field of research. Work has been done on export formats that can be displayed well in the long term. But in many cases, these are individual decisions or scenario-dependent decisions about what you want to achieve and how you want to handle this data or these databases. This is sometimes referred to as "preservation planning." I think this is easy to understand with websites. Websites are also very complex structures. And depending on your own purpose, how you preserve a website can vary greatly.

In some cases, you may only be interested in the visual impression. And it is sufficient to take a screenshot for certain purposes. In other cases, however, you may not only want to be able to highlight and use text, but also maintain interactivity. And then there are completely different requirements.

PM:

Yes, that's right. There are tools like the Wayback Machine for that.

Expert:

Exactly. There are also international consortia that deal with standards for implementing web archiving—that's the key term. And standard archive formats or container formats have been defined, such as WARC [Web ARChive], which is also used by the Wayback Machine, for example. The [formats] are also used by national projects that want to preserve their national internet or certain sub-areas and use tools such as crawlers or storage formats for this purpose. And so the question arises in a similar way with many things. What is the actual preservation goal? And from this, you can then deduce to some extent what measures need to be implemented in order to preserve more complex objects.

PM:

Even though this area is still under development at your organization, do you have any ideas about hardware, such as whether to rent servers or host them yourself? This also involves aspects such as security, access protection, and redundancy.

Expert:

Not all decisions have been made yet. And that's why it's not clear who will be the service provider or what we will do ourselves. But there are, of course, standards, such as those from the BSI, the Federal Office for Information Security, which we will follow. We are active in many areas in this field. There are a few principles that are clear. For example, we will not only have one location where our data is stored,

but these locations must be sufficiently separated from each other geographically so that we achieve independence from copies of the data.

I typically distinguish between three levels of long-term archiving. The first level is where you make sure that the bits are preserved. In other words, that there are no losses, no uncontrolled changes. That's bitstream preservation. You can think of it a bit like media: if a hard drive breaks down, it shouldn't lead to data loss. But it's also a bit more complex than that, because nowadays it's not enough to just have a hard drive; you need a whole chain of technology just so that the operating system can access these bits in the first place. So that's the lowest level, so to speak. The level above that is sometimes referred to by different names, but I would say it's about technical usability. Of course, this involves questions such as: Do I have software that can process this data format? Or do I need additional services? In other words, technical usability. This level is sometimes referred to as "technical preservation" or – perhaps a little misleadingly – "logical preservation." And the third level above that is not specifically digital, but it is also part of it: Can I still use the data in terms of content? Can I still understand it? Do I have sufficient contextual information? These are the three levels that need to be addressed in long-term archiving. And for me, the question of hardware, servers, and media is more at the bitstream preservation level. At least when it comes to the storage level.

PM:

Yes, I can understand that. That means you would also have to provide software that can open the stored data.

Expert:

Exactly, that would be one strategy, at least. A discussion that is no longer really relevant today, but which was held a few years ago – I would almost say in the early days, when people first started to deal with the problem of long-term archiving conceptually – is the question at this level of technical use: Do you migrate files, digital objects from one format to another before you can no longer support one format with software? Or do you emulate old environments and software so that you can continue to operate them without having to touch the object? I would say that this has now become a bit of a sideshow , which is not that relevant very often. But there are, of course, cases that make these questions very vivid. There is the Computer Game Museum in Berlin, for example.

PM:

I know it, I've been there before.

Expert:

Yes, exactly. And for them, it's a very difficult question of how to preserve the usability of old computer games. Emulation plays a big role here, of course. In this respect, it's a rewarding field, because there are many dedicated fans and hobbyists—retrogaming has become a term or field in its own right—who are involved in emulation so that you can have as authentic a gaming experience as possible with old games on new hardware technologies. Looking at your guide, I see that you skipped over the first question a bit, which is who the stakeholders are when it comes to what is stored. I'm almost a little grateful for that, because it's a complicated question.

PM:

I'd like to briefly introduce this. I also asked this question to [the historian from the BGE's knowledge management group], who is accompanying the site selection process from a contemporary historical

perspective. We discussed that there needs to be a kind of gatekeeper who decides what information is worth preserving. He replied that in certain areas, he himself is actually this gatekeeper. In this respect, the question of who the gatekeepers and stakeholders actually are is an interesting one.

Expert:

First, I have to ask you a question in return: What do you mean by stakeholders?

PM:

Well, stakeholder is a very broad term. So anyone who has any influence, any opinion...

Expert:

Nowadays, a lot of people have an opinion.

PM:

That's right. But of course, the public is also a stakeholder. The question is how relevant this stakeholder is to you.

Expert:

Of course it is relevant. The legal wording in Section 38 is not very clear or precise with regard to the objective of long-term documentation, but hopefully the regulation will be a little more precise. But another formulation, which comes from the OECD/NEA research project, which you also mention [in your interview guide], is to enable future generations to make their own decisions. This can be translated into three sub-questions. First: Future generations should know what is in a final repository and how its safety can be assessed. Secondly: Why is it there – what was the process, how did it come about? Traceability. And thirdly: What needs to be considered when accessing this stored waste again or when recovery is necessary? These are some of the questions we use to consider what needs to be stored and how.

And "future generations" is, of course, a very broad term. Of course, the public is a stakeholder, and we don't currently have a fixed classification that I can present to you. What I can say is a little bit about how we have differentiated or differentiate in relation to a long-term archiving system. Of course, there is still room for discussion as to how much influence each party has on the question of what is stored and how. On the one hand, we have the owners of the data. We have "owners of other relevant data," as we call it. This is because some things may not be data, but they may still be important for understanding the data.

PM:

What kind of data could that be? What does that mean?

Expert:

Well, let's say we have borehole data. And that data is considered significant. But in order to use it, I need additional information. And that is this other relevant data. Software might be one example of this. Or contextual information, manuals. We will certainly also have other "memory actors," perhaps something like other libraries or archives that are located in the region, for example, which have also collected data themselves or want to use it. Or people from the fields of media, culture, research, and education. For example, we are now receiving more frequent inquiries from documentary filmmakers

and journalists. And they also have an interest in information. Now they ask us for interviews, and in the future they will probably ask us, "What does the final repository look like?" So these are some preliminary distinctions between actors. We cannot yet expect them to remain unchanged in the long term and to be used for all analyses or decisions. But in the context of a long-term archive, a technical system, they currently play a role.

We have the interested public, who can, for example, submit a request for information. This is open to everyone. There is the Environmental Information Act (Umweltinformationsgesetz/UIG), and you will also ask about the legal basis later. So in this respect, every member of the public is potentially relevant. There are also other stakeholders in responsible roles, perhaps other government agencies in the future, who need information and to whom we must provide information. There are certain external systems, such as data portals, for example. There is the so-called geoportal...

PM:

Yes, every federal state has something like that. I know about the one for Hesse. There you can get City GML models, i.e., building volumes for the whole of Hesse, and other data.

Expert:

There are systems of this kind for which we will probably have to serve as a data source at some point, or from which we may also want to obtain data. The BGE also uses technical systems from which we want to obtain data. These external IT systems are just one issue. And then, of course, we also have various internal players. Under certain circumstances, other departments within BASE may be interested in the content of the data. But other departments are also producers of data themselves. So BASE itself is someone who produces data. We at are ourselves in the role of producers and consumers. And then we have ourselves as a team and IT, which may not have any interest in the content, which decides whether the data is stored. But when it suddenly comes to very high-volume data, questions such as the following naturally become relevant: Do we have enough storage space? How big is our broadband connection? Do we need to expand it somehow? In this respect, there are a whole range of players with whom we interact and where different perspectives exist. But when the question is, "What is archived?", then in our preparatory work on the key points of the regulation, we have simply tried to clarify this somewhat vague term of potentially significant data in order to make it clear that this is the data we are talking about. And to illustrate the objective criteria for selecting the data to be stored.

PM:

We can imagine a scenario when it comes to storing things permanently. How robust is this whole storage process, this decision-making power, in terms of political processes? It is conceivable that someone might say, "Well, we don't really need all this anymore, we can delete everything." Is something like that theoretically conceivable?

Expert:

We are a federal agency and as such have legal responsibilities. If it is the law, then we must implement it. In this respect, the question is: How are laws made? And I cannot speculate here about what is conceivable in this regard. I would just make comparisons. Libraries, archives, and museums always have a certain degree of independence. Archives in particular are allowed to decide for themselves what is kept and what is not. And that is precisely intended as an instrument of democratic control. In other words, those who produce the documents and files should not decide for themselves whether they are

kept or not, because they may have a particular interest. And traditional archives have been designed in such a way that the decision lies with the archive.

Numerous documents and guidelines have been created to this end, in which this is laid down. At a conference, I once heard a colleague from the archive sector say: "We defend democracy" or "Archives are the defense of democracy." I would probably add: "... from a very long-term perspective." But because we are still in the process of setting up, I can't give you any concrete details at this point. However, a very important concept in the field of digital long-term archiving is that of trustworthy digital long-term archives. There are even standards for this. There is a nestor criteria catalog for trustworthy long-term archives [nestor is a German competence network for long-term archiving and long-term availability of digital resources]. There is also an ISO standard. And in addition to the perhaps obvious technical aspects and workflows, it also contains a number of organizational criteria that say: Dear archive, if you want to be considered a trustworthy digital long-term archive, then you must document your financial planning in such a way that the public or whoever is involved can have confidence in your financial sustainability. So these organizational sustainability factors, as I call them, are a major area of responsibility in these standards for trustworthy digital long-term archives.

PM:

Is your work only about archiving and storage, or is it also about creating a kind of information portal? You said that you also expect requests for information. Is there any information at that you would say is stored primarily for the purpose of making it available to others? I see different approaches here.

Expert:

That has not been decided either. As a federal agency, we are governed by laws, and access to the information we have is largely regulated by the Environmental Information Act (Umweltinformationsgesetz/UIG). This means that people can, of course, submit requests, and the Environmental Information Act also contains provisions that require the administration and authorities to proactively provide information. I cannot yet say what form this will take. But in accordance with this legal basis, I would find it entirely plausible to say that information that can be researched via a portal will be made available. However, we must also make a distinction here, because we will certainly also include individual data that contains trade secrets and which we are initially obliged to treat as confidential.

PM:

Personal data, of course.

Expert:

Yes, personal data too. And of course there will be personal data, and it will have to be included to a certain extent. We will have to see how far that goes in each individual case. However, archives usually have special rules for this very purpose when it comes to data protection. In the sense that certain rights under data protection laws do not apply, for example. So I can't go to the Federal Archives and say, "I want you to delete all information in which I am mentioned or delete me from all files." The purpose of an archive is precisely to preserve information. But there are limits, of course. The archive does not have to delete it, but certain types of information sharing may still be restricted. Data protection laws therefore play a role. Copyright and trade secret laws also play a role to a certain extent. And that is the field in which we operate. You also asked explicitly which laws and standards are affected.

PM:

Yes, exactly.

Expert:

I tried to look into this a little more systematically. But it's not that easy to say. So – there is federal and state archiving legislation, there is the Environmental Information Act (Umweltinformationsgesetz, UIG) and then there is the Freedom of Information Act (Informationsfreiheitsgesetz, GeolDG) and data protection laws, which are very similar. Then there is the Geology Data Act (Geologiedatengesetz, GeolDG), and the handling of geological data is a special case. I'd rather not say anything about trade secrets because I don't know exactly where that fits in. But copyright law, the retention periods that companies are subject to... There is the E-Government Act and the Data Use Act, which also state that public institutions are obliged to make data available. So far, we've been looking at the information disclosure side of things, and then of course we also have the data documentation and collection rules, which deal a little more with nuclear facilities. So, in addition to the German Atomic Energy Act (Atomgesetz, AtG) and the Site Selection Act— , a lot of this is likely to be covered in the relevant ordinance—there is also KTA 1404. I don't know if you've come across it before.

PM:

I've heard the term, yes. I don't know the exact content.

Expert:

It's the documentation for the construction and operation of nuclear power plants. I'm not an expert on this type of regulation. But there are also things like the waste flow control guidelines and the Guideline of the Nuclear Waste Management Commission (Entsorgungskommission, ESK) on the Interim Storage and, of course, the internal regulations. These are things where, in particular, what we call data owners are obliged to keep and create documentation. And that is data that we are likely to receive.

PM:

That was also a question I had written down, namely the question of the distinction between storage and transfer obligations. Perhaps you could clarify this: On the one hand, it's about the transfer, but also about parallel or redundant archiving for a certain period of time. So, on the one hand, something is deleted and on the other hand, it is kept on your side?

Expert:

Exactly. The regulation will probably spell this out in more detail; at least, that was our aim in our preparatory work. After a certain period of time, we receive data that is potentially significant. In other words, data on final and interim storage that is or could become significant. We will receive this data from the so-called storage data owners. "Data holders" will be defined in the regulation. This will not be just any actor, but they will be defined on the basis of other laws. But of course, for example, the Federal Company for Final Repository (BGE) and the Federal Company for Interim Storage (BGZ). These are two very large [actors]. And when the data is complete, for example, it must be offered to us and we will take it over. Until then, the owners of data may not delete this data due to long-term documentation requirements. Once we have received the data and confirmed receipt, we will of course also carry out integrity tests and quality control to verify that we have received the data correctly and can take responsibility for it.

From that moment on, we no longer have any objection to the deletion. But it may of course be that the owner of the stored data still has to keep the data for other reasons – tax documents or whatever. We cannot judge that. And we cannot take that responsibility away from the owners of the stored data. That is not legally possible. So we cannot, so to speak, be the archiving service for the owners of data, because that would mean we would also assume responsibility for these other legal obligations or contractual obligations that they have entered into. But it is conceivable that the data would ultimately only be stored for us, we would then take it over and say, "Now you can delete it." And then, as the owners of the data, they would delete it on their side and we would then be the sole source of the data.

PM:

What about the timeliness of the data? Once we have a final repository for the highly radioactive waste, will you be supplied with data continuously or in packages? How should we imagine this?

Expert:

In the discussion for the regulation, there are annual deadlines on which we receive lists of lists and, based on these, data is then sent to us. It is also possible that there will be updates. And some data is also updated...

PM:

Exactly, that's the question. If you have an existing system where you expand things, but the file still has the same name and has only become a little bigger.

Expert:

Exactly. So in a few places it really depends on the regulation, but one consideration would be what data should be offered in these annual overview lists. One possibility would be that it is completed data that was finalized last year. And then it would be impossible per se for it to be updated again, one would think. On the other hand, updates are sometimes made retrospectively. Conversely, one requirement, as we would currently formulate it, is the traceability of the process. This means that interim versions used for a specific published report or for decisions are also important in order to understand what the status was at that point in time.

PM:

So a kind of versioning as in classic software development?

Expert:

Exactly, there will have to be versioning. That much is clear. But I can't give you any details at this point, especially since this regulation is not yet in place. And it should specify more precisely when what is to be offered.

PM:

OK. We have already discussed storage and transfer obligations. There was a research project [Long-term durability of paper or "Labest Papier"] that looked at the long-term durability of paper. And the target was set at 500 years. You've already said that long-term archiving means permanent archiving. What interests me here is: How did you arrive at 500 years when the goal is actually to store data for much longer?

Expert:

There is a specified date. This means that radioactive waste should be storable for 500 years. And that is at least a clearly identifiable milestone where you can say: OK, we can definitely test that. There are other factors to consider, but if we assume a worst-case scenario where no organization takes care of the documents, then the paper should at least survive those 500 years. When you read about the 500 years—I can't give you the exact paragraph in the Federal Site Selection Act ( ) —the reason is usually that there is a paragraph that states that retrievability must be guaranteed for 500 years. The final repository should be planned in such a way that the decision we make now – as best we can – can also be revised by future generations. And that includes the retrievability of the waste after 500 years.

PM:

The advantage of paper is that I can lock it in a room and leave it there for 500 years. With digital data, you need an infrastructure that also needs to be maintained. Well, you could maybe put everything on a hard drive and see how long it lasts.

Expert:

Certainly not 500 years. Now it gets a little complicated. Because we have the task of preserving information. And the primary question is not whether to use analog or digital. But of course, most of all information today is created digitally. In the phase we are currently in, this information is being created. We collect it and try to store it, make it accessible, maintain its usability, and make it available again. How long do we do this? That depends on many factors. Ideally, I would say forever. And then normal operation would simply involve updating systems, media, hardware, etc., of course. And that way, you can maintain usability in the long term by simply having an institution that takes care of it. In this OECD/NEA project "RK&M" – even if it doesn't originally come from this project – a distinction is made between different time frames, namely short term, medium term, and long term.

And short term is – which may sound surprising to the general public without this technical context – until the final repository is closed. That is a very long period of time by our standards, but in this context it is defined as short term. Medium term – I don't remember the exact wording – essentially means as long as an institution takes care of it and supervises it. And long term is the period when the final repository is no longer supervised. I believe that is the wording. And I think it is reasonably plausible to say that if the final repository is no longer supervised, then the archive may also no longer be supervised. And that would mean that a digital archive works in the medium term because there is still an institution that takes care of it, but not in the long term. What do you do then? To a large extent, this is covered in the research project you also cited, which involves having a variety of measures in place to enable reconstructability and traceability. That is what we have to rely on. I think it's a totally plausible assumption to say: We have to work digitally for now, and that's how we'll be working in the near future. And at various points in time, we'll transfer part of our digital information to particularly long-lasting media or possibly just to particularly robust paper. You also mentioned that we have research projects on the long-term durability of paper. And with paper, you can bridge a few thousand years.

PM:

Yes, I think papyrus lasts a while.

Expert:

Exactly, it depends on many factors. My personal opinion: Given where we are right now, we don't have to make a decision about media today. But we are researching it because we at BASE, or rather our

successors who will be responsible for this, will have to make a decision on media at some point. And that's why we are conducting research for this purpose, so that we can eventually select the best possible media and the best possible storage method – coding also plays a role here – based on the state of the art in science and technology at that time.

PM:

Now, in this table [from the final report of the OECD/NEA research project "Preservation of Records, Knowledge and Memory (RK&M)], long-term archiving is only described as one possibility. I am surprised – please correct me if I am wrong – that these other approaches [such as markers and education] sound softer. It seems to me that there is no legal basis or obligation for them yet. In summary, in my opinion, this regulatory framework seems to focus heavily on data storage and archiving, but less on these other options.

Expert:

To my knowledge, we do not have any legislation on markers in Germany. But that may change at some point. This is also something that we see from a technical perspective, that there is a need for action. And I believe it is something that will need to be discussed in the coming years. In general, I believe that the important value of this RK&M report is the general strategy of saying that it is not enough to have one mechanism, one approach. Instead, you need several that are interlinked, because over such a long period of time with unpredictable events, the best way to ensure that information is preserved is to have many things that reinforce each other. In other words, you need a network of measures that are interlinked.

PM:

The Swiss ordinance [ENSI-G09 "Guideline for Swiss Nuclear Installations – Construction and Operating Documentation"] dealt specifically with operating documentation for nuclear installations. You have already mentioned KTA 1404, which would be the German equivalent...

Expert:

...Perhaps.

PM:

Yes, OK, I'll check the content again. There [in ENSI-G09], the aspect of storage and transfer obligations was also addressed. PDF-A is mentioned as the preferred document format. PDFs are also ubiquitous in current document management systems. Do you think that the PDF standard is well suited for long-term archiving?

Expert:

I am familiar with the discussion among the long-term archiving community on the subject of PDF. It is also a container format. It can hide a great deal. And PDF-A is definitely a step forward. That's how I would put it. That doesn't mean that it solves the long-term archiving task at the format level. It's a step forward in many areas, but there are also sub-versions. However, I can't comment on that in detail at the moment. It's quite possible that we will use this [format] in many cases. It's impossible to say for sure. There are different versions of PDF-A, and even in PDF-A, you can't do some things very well. If you're really interested in this, I would refer you to the experts at TIB Hannover, the Technical

Information Library in Hannover. We have a number of people there who really look into the PDF with the hex editor...

PM:

Interesting, but let's stick with the guideline [ENSI-G09] for now. It refers back to ISO 14721...

Expert:

...and the OAIS model [Reference model for an open archival information system].

PM:

Exactly. I was familiar with this model, at least by name, from our basic lectures in computer science in construction. Until now, it seemed to me like a relic from decades ago, but it seems to have a certain relevance after all.

Expert:

I have to interrupt briefly. You were familiar with the OAIS model from your studies?

PM:

Yes, in the context of information processing and human-machine interaction. The model was mentioned in the introduction.

Expert:

Because as a reference model, it is one of the central models in the field of long-term archiving.

PM:

Figure 4-12 [Information Object Taxonomy] of ISO 14721 shows that it is not just the "information object" that exists, but that further information on interpretation and representation must also be included as components. Earlier, we also talked about databases and management systems. Is this something you look at systematically? Do you really take the trouble to consider the pillars underlying each information object in advance?

Expert:

A comment on OAIS: This is, of course, a reference model. Ultimately, it is there for comparison and provides a good portion of the terminology used in the professional community of long-term archiving. But it is not an implementation standard. Separate standards have been created for this purpose, such as metadata standards. More specifically: how best to create them and what to use, how to record them. So yes, of course, in the language of the OAIS, representation information is needed. Under certain circumstances, it is sufficient to specify the format, i.e., to say that it is a specific PDF-A version. There are other cases where more information is needed. And for this, we naturally fall back on standards that exist in this international community—but again, we are still in the process of development and do not have a finished standard. One important implementation standard is PREMIS [Preservation Metadata Information Standard] from the Library of Congress. This is closer to implementation. So these questions are relevant. The specific implementation tools have not yet been decided. To a certain extent, this also depends on how the long-term archiving system we are purchasing works. But they all support PREMIS.

PM:

My question was also aimed at whether those who provide you with the information also have to ensure that they provide these lower-level building blocks. For example, that they receive a note indicating which software can be used to open a particular file. I mean, it makes a difference whether I just receive a container with file types that I may not be familiar with, or whether I also receive information about which software can be used to read them.

Expert:

It has not been decided how much of the data we can request. But of course, there is a need for much of this information, i.e., contextual information of both a content-related and technical nature.

PM:

Yes, I stumbled across Figure A-1 "Composite of Functional Entities" from ISO 14721. Then I saw how extensive the whole thing is...

Expert:

As I said, it's a reference model. That doesn't mean that you fill in every box with a single person. And some things are machine processes. It helps to think of many places in these large, gray box categories as ingest, data management, archival storage, access, preservation planning, and administration.

How the individual functions are implemented varies greatly. For communication within the company, I don't use the OAIS at all, but generally try to explain what memory organizations do. Namely, they take on data – which is a bit like ingest here. You then have to index the data, so you have to know: What is it? This is hinted at a bit here in data management. You have to preserve the data. That's probably archival storage and preservation planning. And well, if it's just there but no one can access it, then it wouldn't make any sense. So you also need to have some kind of access or mediation. That's access. So in these four basic categories, it definitely helps to think about it and then make it more precise step by step.

PM:

OK. We've already talked a little about the next point: migration. The standard also provides a few recommendations for action. Can you think of anything else?

Expert:

Well, further concepts have of course been developed in this long-term archiving community. The OAIS also has its own terminology, which may not be the most widely used terminology at present. One concept is that of "significant properties." What are the essential properties that need to be preserved? And that usually depends on the target group or the target usage concept. Here's an example that's not from our field, but is still easy to understand: At the Computer Game Museum, interactivity is obviously a "significant property." There are other cases where perhaps only the visual impression is decisive. And that's why it's something you have to think about a lot when considering how to preserve things. What are the processes that I want to support later on, that need to be possible? What are the target groups? "Designated communities" and "significant properties" are key concepts in long-term archiving theory.

PM:

Are there any models in other European countries for all these aspects we have just discussed? I am broadly familiar with the technical concepts of interim storage and final repository in Switzerland and France, for example, but I don't know whether there is a structure like BASE there. Do you know, and is there an exchange of experiences and best practice examples? They are facing exactly the same challenges.

Expert:

When it comes to the long-term archiving of digital data / long-term preservation, I would say that there are no best practice examples, because that might be asking too much. "Best" already implies that there are several, and that is then the best in best practice. There is simply practice [in the sense of practical application] and research. And these institutions—the counterparts in France and Switzerland, ANDRA and Nagra—with whom we work—not me, but a colleague of mine—in OECD/NEA committees. There is not only this RK&M project, but also other working groups in which we collaborate with counterparts or reasonably comparable institutions from abroad and discuss such issues. However, this remains a current topic of research and discussion.

PM:

What is the status of these institutions and how far along is BASE?

Expert:

[We are in a good position.] Last week, I attended the iPRES conference [International Conference on Digital Preservation], where a project partner from ANDRA presented the results of a research project on how they record digital data in coded form on paper. This concerned the storage medium of paper, but also how it can be used efficiently for digital items. And I would say: Yes, different countries have different priorities, but essentially they are at a similar stage. Some already have their own archiving institutions. In the UK, for example, there is Nucleus, in the far north of Scotland. They have their own building with storage facilities. Many other countries are not yet at the stage where they can say they have their own archiving institution. So, I would say that the status varies. No one is finished, no one has a complete concept, but related issues are being worked on internationally. And people are also exchanging ideas.

PM:

Is there an exchange beyond Europe?

Expert:

Yes, but I'm not the expert on that. I can't tell you what the situation is in the US with regard to information preservation for final repository. Many of these highly publicized and imaginative ideas originally came from a project in the US that focused on how to inform future generations that they should not dig for raw materials here...

PM:

...yes, nuclear semiotics and putting up pillars in the landscape.

Expert:

Exactly. Much of this comes from a project in the USA. But I'm not an expert on this.

Expert:

But my impression is that, simply because many countries in Europe that use nuclear power are located in a small area, interaction and research on this topic is perhaps most advanced here at the moment.

PM:

It's just a natural fit.

Expert:

Yes, exactly. Yes, there are also people and institutions in other countries and continents that are working on this. I am mainly familiar with European countries at the moment—perhaps it's a coincidence, but I suspect it also has a little to do with the focus and these framework conditions.

PM:

OK. Then we're already at the third complex [of the interview guide]. This initially concerns a passage from the research report on the Labest Papier research project. It was pointed out there that there is a kind of triad for long-term archiving, consisting of paper, digital electronic, and optical archiving. And it was pointed out that this is reflected in many standards. We always hold the single source of truth principle in high regard when it comes to working with databases. We also strive to apply this principle to BIM. In the research project in question, however, the focus is on redundancy, which also has its advantages. Can you discuss how these two very different concepts can be applied to long-term archiving?

Expert:

I can't speak for this passage and the research project. I don't know enough about it. But one important difference I see between the case you presented and the case in this more archival context is, of course, that in the archival context, the data is closed. , it seems particularly important to me to ensure that no inconsistencies arise. I believe that this risk is no longer seen in this area of archiving, because no one should change anything in the data. That would indicate a serious problem, or it would have to be made clear that a change had been detected here. We fulfill our function by checking for such problems and changes and becoming aware of them. And the aspect of preservation is much stronger when you say, "OK, if we only rely on one means, then there may be systematic problems that jeopardize the integrity of the information."

With a single source of truth, you could also say that, in the strictest sense, the information only needs to be stored in one place because if you have mirrored the database in two places, the question arises as to whom you should believe in the event of an inconsistency. The thinking here is that this inconsistency cannot actually arise in regular processing because the data is usually considered complete. And that is why planned redundancy across different types of storage, different locations, and possibly even different organizations is important. For bitstream preservation, this first level, the basic principle can actually be formulated very simply: it is about creating as many independent copies as possible, which are then checked regularly and restored as quickly as possible in the event of discrepancies or damage. That is actually the whole principle. But this "as independent as possible" means technologically, organizationally, and geographically. There are a whole range of types of independence involved. And we know from the field of technology that if you only use hard drives from one production batch, it is quite possible that they will have a common fault or a common characteristic, which will cause them to fail at the same time or within a sufficiently similar period of time.

PM:

Do you think this triad of different media will really be practicable in the future? Well, if you have analog data, you can digitize it and turn it into a PDF. But if I start with digital data—I'm thinking of BIM, for example, an IFC file that consists of ASCII code but can only be interpreted by humans to a limited extent—then the question arises as to whether these files can also be printed out. If I start with CAD planning data, I can of course import it into an editor. But does it make sense to print this content just to have the information on paper?

Expert:

I wouldn't say that it has to be this triad. But you really have to look at the specific scenario in question. I remember visiting a data center a long time ago where bank data was also stored. And it was clear that they also wanted to have optical media. It was important because it allowed them to guarantee immutability. Unlike with a hard drive, where someone can cast doubt on it in court. That's much more difficult with optical storage media. That's a different scenario. So the right choice of media depends very much on the case and the scenario. I don't want to make a general statement about long-term archiving as it is formulated here. But: Independence from copies at very different levels is an important factor. Targeted redundancy.

PM:

We have already talked a little about reading or decoding devices. If I have paper, I don't need anything else to read it except this information carrier; the human eye does that. It's different with digital data. I don't know if this needs to be specified in the regulation, but how much responsibility do you also have for the possibility of reading?

Expert:

We take responsibility for the information and its usability. As I understand it, this regulation will not primarily concern physical objects, but rather content. And the question of reading and decoding devices is more a matter of the terminology used earlier for medium term and long term. So as long as we have an institution that takes care of this—us or a service provider—and ensures that we can continue to access the information. Take hard drives, for example: as long as you are in an area where you can constantly take care of them and replace and exchange the hard drives, there is no reason not to use hard drives as a storage medium. Only when we say, "OK, we now have a level of information that we want to secure in such a way that even if it is neglected in 50 or 100 years, someone will still be able to access it" – then the storage medium question becomes relevant.

PM:

That would be long term, yes.

Expert:

That is specifically long term. To a certain extent, the difficulty lies in the fact that sometimes you don't know when an organization will stop taking care of certain things, because that depends on the circumstances. I know this from the field of research data: At the beginning of Donald Trump's first presidency in the US—I can't say exactly which data centers—there were, to my knowledge, research data centers that no longer received funding and therefore certain data sets could not be preserved and continued and were lost to research. In a sense, this is a normal process in human history, that not all data is preserved. Some people find this regrettable, but on the other hand, it exceeds the capacity to

preserve everything. But as long as we have this controlled process and an organizational unit, reading or decoding devices are a secondary issue.

PM:

So you mean that as long as the organization is there, with staff and perhaps also active research to keep it alive, it is a sufficiently stable framework.

Expert:

It is possible that in the near future, an additional copy will be stored on particularly durable media or something similar. But in terms of short term, medium term, and long term, strictly speaking, it is only necessary in the long term to ensure that it has the highest possible degree of independence. Of course, there may be other factors. In some respects, it may be more cost-effective to dispense with reading devices or decoding devices. There are discussions about so-called cold storage. This also affects the large so-called hyperscalers, the large providers of data centers, such as Amazon, Google, and others. The have huge data stocks, but most of them are accessed extremely rarely. The vast majority of accesses are to what is currently very topical on social media and elsewhere, but this is only a very small part of the data. And they always try to store the information on the best storage medium for this frequency of access.

They accept that accessing old data takes longer because it is significantly cheaper. In extreme cases, they would use tape storage for data that is requested extremely rarely and where they think it is justifiable. And the data that is requested most frequently is stored on SSDs or even in RAM. This is a hierarchy of storage. And for cold storage, research is continuously being conducted into new media. Things like DNA, ceramics, and so on are being considered. And then you always ask yourself the question: What are the reading or decoding devices that I need or don't need for this? And these are questions where, in a sense, the main criterion is whether it can establish itself on the market. That's where economic issues become relevant.

PM:

The research report mentioned that from 2023 to 2025 there will also be a research project specifically on the long-term durability of digital media. I found that very interesting, and it fits in with what we have just discussed. In principle, it is the logical continuation of the investigation into the long-term durability of paper. Do you know more about the scope of this project and the methods used there?

Expert:

I am not supervising the project myself, but a colleague is. The project looks at the durability of media under various factors. It also distinguishes between which media are particularly suitable in cases where they are actively maintained or where neglect occurs. And since we don't have to decide on media for the next few years based on this project, I think it's more about getting up-to-date info so we can compare how the field develops in future projects. Over the years, this will help build up an overall picture of how storage media is developing and how to choose it. There is also a methodological component to how media is evaluated. I cannot give you any concrete initial results.

PM:

OK. Then we'll just have to wait until the end of the project. We've been talking about information. In knowledge management, there is a kind of pyramid. At the bottom we have data, in the middle we have information, and above that we have knowledge. In my concept, knowledge is always the linchpin. I

assume that it makes sense for nuclear technology not only to have resources at the document level, where the focus is primarily on data and information, but also to focus more on this knowledge component. How relevant is this for you in long-term archiving? Even the law "only" refers to information.

Expert:

The colleague who is in charge of the media project is also responsible for knowledge management. This is relevant for us as an institution and also in long-term archiving. I would say that this topic actually takes precedence over long-term archiving or, to put it in terms of the OAIS, is super . The question is: What is the relevant information? But of course, to a certain extent, it also has to do with the fact that the information must remain interpretable, so contextual knowledge is needed. In principle, the OAIS also provides a theoretical tool for describing contextual knowledge: "representation information." Contextual knowledge is not only technical, but can also be semantic, to use that as a simple umbrella term for contextual knowledge in terms of content. But that is something we are still working on. We need to develop indexing concepts that make the information navigable. Without having the information available, we cannot yet present a finished product or make any assumptions. Of course, we have a certain system, for example, through the published draft list in the key points for a regulation, which is definitely significant data. There is a certain logic to how this is done, how it is structured, categorized, or how classifications or ontologies may be developed.

We are looking at things, but we cannot decide in advance. There is a kind of ontology draft from an OECD/NEA project – I believe it is called RepMet [Radioactive Waste Repository Metadata Management]. Entity classes have been formed, and I would say: Yes, we need to move in this direction, building an information model of the final repository system together with the waste, the actors, and the location, and using this to make the information navigable and interrelate it. I also found it interesting in your presentation that you are also working on an ontology. I think that's the direction we need to go in.

PM:

There are different levels. The simplest would be a taxonomy, which is something that document management systems work with, i.e., typical keyword indexing. And ontology would then be...

Expert:

...logical conclusions.

PM:

Exactly. Reasoning afterwards – that's a big topic for us right now. As you've seen at IFC, there are individual classes, such as IFCDoor, IFCWall, and so on. That's already a kind of taxonomy. There are research projects that aim to find "the" ontology for construction or environmental engineering, or that strive to combine ontologies for individual sub-areas. But the question is, of course: if I want people to be able to use the archived knowledge in 500 years' time, this contextual knowledge is needed. How much importance should be attached to technologies such as big data or AI in advance, for example for classification? One could also say that people in 500 years' time should solve this themselves and that we today are only responsible for the information.

Expert:

Yes, I believe that different time phases have different tasks. The tasks we have now: Yes, we are also looking at these topics of ontologies, AI, and big data. But more so for our current tasks. But then putting the information into a form that, even if there are organizational discontinuities, local or regional disasters—whatever—it is still understandable after a break in care or stewardship? That is a special task . I wouldn't put that in one category. So, of course, we will now look at how we can possibly use AI methods to simplify indexing or improve the ability to find information. But that is different from what will be prepared in the future so that future generations can use the information as unconditionally as possible. You have to distinguish between the two. The RK&M report you quoted also contains a few different tools. Among other things, it mentions that separate document sets or documents are created, such as the Key Information File or Set of Essential Records.

Separate products are created for specific purposes. And perhaps AI methods are used for this. But we also need them beforehand for other tasks. However, in this RepMet initiative, entities are defined and related to each other, which should be recorded in the area of final repository for information. In my opinion, it is something that must be continued. It is not the last word on the matter. But that is how far this initiative has come. And it is always nice to be able to refer to a standard. It's just that there are often too many standards. And we have to see how we can take these up and continue working with them.

PM:

Let's come back to my concept. I told you that I am currently working with a graph database. My initial concept was based on bringing the Knowledge elements as close as possible to the BIM model and treating it as a single unit. I have since moved away from this and now believe that it is more future-proof to keep them separate. You have the BIM model and, of course, each Knowledge element can refer to individual elements within it. In BIM models, each object has a unique ID. You can use this to make a cross-reference. This means that not everything is contained in the BIM, but the actual knowledge is stored in the graph database in the form of individual data records. Is this approach promising in terms of possible long-term archiving? Does it make sense to keep the "modules" separate and not have one large "block"?

Expert:

At the risk of using a cliché or a phrase: it is better not to have monoliths. And this approach to integration sounds to me like it is a large model, which would probably be more difficult to maintain than having the BIM model and a variety of tools, each of which can be traced back to individual standards. And then I have a separate graph database here, or something that hopefully is also based on standards, or at least incorporates aspects of them, because standards are proprietary. So I believe that modularity is an advantage in terms of maintenance and long-term archiving.

PM:

I'll gladly take that on board, because it supports my concept. Yes, great. Then we've now come to the end of my guide. Are there any other topics you'd like to talk about?

Expert:

Probably many. You are concerned with knowledge management. And perhaps the biggest challenge or risk you can have in relation to digital technology is not, in my opinion, when it comes in the form of a clearly definable digital object such as a file or similar, but rather when information and knowledge are contained in software or a service. I touched on this a little at the beginning of when discussing the

distinction between a hardware object and the software used to control it. In other words, where the knowledge is embodied—not simply in a document, but in software or services. I believe this is where a particular challenge lies, and we will encounter these cases.

I know that this has not yet been formulated in the form of a guiding principle that can be taken home, but a large part of the issues in long-term archiving must actually evolve towards determining the service, benefit, and purpose behind something digital. This is not currently common practice in the long-term archiving community. But from these questions – do I now take this or that preservation measure for this individual data object? – I actually have to move more toward overall usability. What is the service that is actually being used here? Your concept involves a chain of elements that together form a network so that someone can use something, right?

PM:

Yes, exactly.

Expert:

You have – sometimes people call it a technology stack. But it's actually more than a technology stack; you actually have a whole network of dependencies. So you have to use the entire network to access something. And then you are no longer dealing with the preservation of a single object, but of a larger network. And that seems to me to be a particular challenge. And that is what makes individual solutions so fragile, because then there is only this one access point to use this object; in other words, this one single path through the dependency graph. So the dependencies form a graph, and—this is an almost forgotten theory in the long-term archiving community—to evaluate the preservability in such a dependency graph, you look at so-called view paths: How can I access an object? That is a path in this dependency network. Actually, you have to make sure that you optimize this network so that you have as many paths in it as possible. And this is not particularly difficult with typical digital objects, but it becomes difficult when what I want to preserve is actually the service.

Several years ago, there was an excellent keynote speech at the iPRES conference, which must have been in 2008 or 2009: "How to preserve virtual worlds." And then it struck me: Oh, that's really difficult. For one thing, the software doesn't run locally; it runs somewhere on a server. And then, even worse, virtual worlds and virtual computer games thrive on the fact that other people are in them. In fact, what you need in these cases is something like historiography, so that you can really preserve something. That's going a bit too far, but in many cases you have to think more in terms of preserving an ecosystem...

PM:

Yes, "ecosystem" is a very good way of putting it.

Expert:

... and as many viable paths as possible in a network. It's a bit like that with "solutions" in the sense of special applications and interactive websites that are backed by a service. If I want to preserve the solution, I actually have to preserve this service. I have to obtain the software, which is often custom-made. And then I'm very keen to obtain the software, to obtain a whole little technology zoo or ecosystem. That seems to me to be the biggest challenge. All cases in which knowledge can be compressed and made explicit, so that it is contained in a document or something similar and does not have such dependencies – those are the good cases. But since we also mentioned databases: databases

are very often connected to such services. That's where I see the challenge. It can involve a great deal of effort.

PM:

Yes, I'll definitely take that on board. When I think about my concept now, I naturally have a graph database that contains the individual pieces of information. But of course, the composition takes place in a game engine and visualization is done with a VR/AR headset. I would have to see whether it would be possible to reconstruct the relevant information based solely on the database if these glasses were no longer available at some point or if the software or game engine, etc., changed. So the question also arises: Is a module not only easy to integrate, but does it also contain information and knowledge on its own?

Expert:

Yes, and I mean, you use Unity. That's good, of course, because it's probably the most widely used game engine. And that's also a sustainability factor, of course, that we use something that is currently widespread and therefore promising. But the same applies to both media and software: the market position and cost-effectiveness of these things are an important factor in determining whether they can be used in the long term. This question of media is a good example, because every few years there are reports about "the super storage medium." Science journalists report on it, and it is indeed very interesting and inspires the imagination. But these technologies almost never become widespread and are almost never efficiently usable. Why? Because they cannot establish themselves on the market. Because they would theoretically be better in individual cases, but: Any new storage media technology that wants to assert itself today must be used by the big hyperscalers. Otherwise, development and mass production are simply not profitable. Then other things are more efficient and it cannot establish itself. And as a result, it does not last because the technology paths shrink too much, become fragile, and eventually break. I mentioned earlier the organizational factors for trustworthy long-term digital archives—these must be taken into account.

PM:

Yes, I find that an interesting point, that you can simply rely on what has already established itself on the market with a clear conscience and not always pin your hopes on technologies that are not yet available.

Expert:

This is something where you have to be modest and aware: the market is also changing, and you have to keep track of the different types of technologies so that they remain usable. And for the time when we receive the digital data and actively work with it, we have no other choice. So, we have to do this because this is how the data is created, this is how we get it, and we just have to pass it on.

PM:

Yes, OK. That was a good closing remark.

Expert:

I hope there was something interesting in there for you.

PM:

Absolutely. Thank you very much for talking to us.