“Catching up with the present: Archiving born-digital records of architecture and design”

Tim Walsh

Digital Archivist, Canadian Centre for Architecture

19 April 2016

SLIDE 1

Hello. My name is Tim Walsh. I am the Digital Archivist at the Canadian Centre for Architecture – or CCA – in Montreal. Yesterday I spoke of CCA’s experiences with the Archaeology of the Digital project, and used a few examples of archives from the projects collected to give a sense of the work involved in making legacy born-digital design records accessible for research. Today I’m going to approach the issue of CAD preservation and access in a slightly different way, with an audience of archivists in mind.

SLIDE 2

A quick overview of where this talk is headed: I’m going to start with a brief introduction to CCA and our born-digital archives, to give a bit of context for those of you who weren’t able to join us yesterday. Next I will very briefly follow up on Julie’s talk with some ideas of why born-digital design records are significant and worth preserving; talk a bit about digital preservation standards and practices and their relation to computer-aided design; give something of a detailed overview of the archival processing workflow as it currently exists for born-digital material at CCA; and finally end on what is hopefully an optimistic note, by sketching out the landscape of the different people and projects working on this same issue.

SLIDE 3

So let’s jump right into the introduction. For the sake of those of you who were here yesterday, I’ve mixed up the pictures on some of these slides, so hopefully that will make up somewhat for the repetition.

SLIDES 4-10

What is CCA? CCA is a Research Institution, where exhibitions, print and election publications, a bookstore, public programs and the Collection all equally are part of the investigation on architecture being a public concern.

CCA’s activities and collection are not intended to capture our build environment, but rather to explore the architectural discourse, experiments, with an international and theoretical perspective. In figures, CCA has:

- 2 main exhibitions, 3 smaller octagonal exhibitions, and 3 hall case presentations per year;
- About 50 lectures, seminars, and public conversations per year;
- A collection including approximately 190 archives, 100.000 prints and drawings, 60.000 photographs, prints, and albums; 240.000 monographs and 5000 series of periodicals; and
- Approximately 80 full-time employees

SLIDE 11

Since 2012, CCA has been engaged, among other pursuits, in a multipart project curated by Greg Lynn entitled Archaeology of the Digital. The project is comprised of in-depth research into digital architecture and a historical reading of the trajectory of digital architecture through twenty-five key projects, from early experiments in the 1980s through to work in the

2000s. These projects were developed by some of the protagonists central to the debate on architecture and digital technology, and each has influenced recent architectural history in a particular way. The research has resulted in a new acquisition strategy for born-digital material and the formation of a digital archive. Since 2013 CCA has presented two Archaeology of the Digital exhibitions, with a third openly shortly on 10 May 2016. After closing at CCA, each of the shows has traveled to Yale. The CCA has also produced one print publication already, with another forthcoming in 2016, and is continuing work on a series of digital publications on each of the projects, incorporating screenshots, videos, and original born-digital files from the archives alongside transcripts of interviews between Greg Lynn and the participating architects.

SLIDE 12

Here you see a listing of the Archaeology of the Digital projects. The archives of each project on the list (with the exception of Frank Gehry's Lewis Residence) and occasionally other related projects or even a firm's full fonds have been accessioned into the CCA Collection. These archives contain diverse records on an equally diverse range of mediums and formats, but what is perhaps of the greatest interest for this talk is that every single one has a born-digital component comprised at least in part of original CAD or 3D modeling files.

SLIDE 13

Of course, while CCA's digital preservation and born-digital archiving programs were largely born out of the Archaeology of the Digital project, we have also acquired other born-digital materials in the years since Archaeology of the Digital began. Nearly all of CCA's recent archival acquisitions, including the archives of Ábalos & Herreros, Bijoy Jain, Brian Boigon, Álvaro Siza, and Günter Günschel, have a born-digital component. In many cases, the born-digital component of these archives is substantial, containing drawings and primary documents not duplicated in the paper records of the archive.

SLIDE 14

To take a step back for a brief moment, it might be instructive to think about just what it is we are trying to document and preserve, with architectural records in general and born-digital in particular.

SLIDE 15-17

First: the built environment. As a rule, CCA is not overly concerned with the built environment – our collecting scope is international, and focused more on architectural process and research than its final, built products. Yet the importance of preserving the built environment is not lost on CCA or its founder, Phyllis Lambert.

Here you see the Shaughnessy House, a federally and provincially recognized historical site, and one of the few 19th century residences open to the public in Montreal. The house, built in 1874, was purchased by Phyllis Lambert a century later -- a time when much of the existing built environment was being demolished to make room for highway on-ramps and modern high rises – and carefully restored over a period of years. The modern CCA building, designed by Peter Rose and Phyllis Lambert and constructed between 1985 and 1989, surrounds the Shaughnessy House on three sides and provides a public access to the building via the main gallery spaces.

There were once many residences like the Shaughnessy House in our section of Montreal. Aerial photos demonstrate that this is very much no longer the case, with the (**SLIDE 16**) aforementioned highway on-ramps on one side of the building and (**SLIDE 17**) modern high-rises on the other.

While the Shaughnessy House is lovely, it is very much an exception. Our built environment changes frequently, often leaving only archives and records behind to document what was once there.

SLIDE 18

Another element of architectural practice to document and preserve, and one more common to contemporary CCA's curatorial interests, is the design process itself: how an idea goes from an idea to a sketch on a napkin and becomes iteratively refined until it reaches its final form. Here the widespread use of 3D modeling technology abruptly and irreversibly changed centuries-old practice: for instance, plans, sections, and elevations may no longer be the primary means by which an architect designs a structure, but instead outputs from a three-dimensional model that can be viewed, manipulated, and queried in previously undreamt ways.

SLIDE 19

We can also ask questions of architecture's engagement with larger social trends and issues: how architecture both responds to and informs questions of environmentalism and sustainability, housing, immigration, and other social and political issues are important questions, and ones that are routinely addressed in CCA shows and publications.

SLIDE 20

And, finally, we can seek to document significant changes in the practice and education of the field as a whole, such as the "digital turn" following widespread adoption of computer-aided design and other computational tools. This is one of the aims of the Archaeology of the Digital project: to better understand the ways in which digital technologies have changed or been used to fundamentally change the practice, study, or even understanding of architecture.

This is perhaps the most straightforward example of the importance of saving born-digtial design records, but the fact is that a researcher or member of the public who is seeking information about any one of these research topics for any work done since the 1990s or so will likely need to consult born-digital records. If we want to continue to document and interrogate contemporary architectural practice, we have to be able to accept and steward the types of materials that architects produce.

SLIDE 21

Which brings us to the question of how we actually do that. My aim for the next few minutes is to paint in broad strokes a picture of the landscape we're inheriting, both in terms of computer-aided design and digital preservation practice, and explore a bit the intersections between the two.

SLIDE 22

Perhaps it's best to start with some simple definitions and common ground. When we talk about computer-aided design, we mean at its core a collection of lines in 2D space, or lines, surfaces, and solids in 3D space. Of course, in reality CAD software are much more sophisticated, with layers, rendering and animation abilities, clash detection and other computational features, and so on. With Building Information Modeling, we see the introduction of the idea of total lifecycle management – a project can be designed, engineered, priced, and built all from the same file, allowing all involved to work from the same data and annotate as needed. BIM software like Revit is rapidly becoming the new norm, even if the collecting delay means that it hasn't yet entered into the collections of institutions like CCA.

Across the board, with CAD, 3D modeling, and BIM, we are talking about varied and highly complex software that is typically also highly proprietary. These features and others have given this material a reputation in the digital preservation community as a notoriously difficult outside case. Alex Ball's 2013 report "Preserving Computer Aided Design" for the UK's Digital Preservation Coalition provides an excellent analysis of why this is the case, as well as an overview of the pilot and research projects conducted to that date.

We can also say that CAD is multi-industry. Architects and engineers use CAD, but also animators, filmmakers, and video game production companies. Much of the software used by cutting-edge architects in the 1990s, such as Alias and Maya, was predominantly developed for and used in the film industry, and only later adopted by architects seeking to push the envelope.

I think it's also safe to say that we seem to be at an interesting moment in the development of these technologies: at the same time that the field appears to be standardizing toward a small number of BIM software such as Revit that themselves were developed to promote standardization and interoperability, we are also seeing an increase in the use of scripting, as evidenced by the widespread interest in and use of Grasshopper and now Dynamo. What tensions exist between these two trends – and their implications for data archiving – remain to be seen.

SLIDE 23

When we talk about the wide variety of CAD software and file formats, what exactly do we mean? This is an admittedly partial list that maybe starts to illustrate the situation. With the exceptions of Revit, Digital Project, SketchUp, and ARCHICAD, all of the above can be found in the Archaeology of the Digital digital archives. For each of the most common formats, such as AutoCAD's DWG, Microstation's DGN, Rhino's 3DM, Maya's MB, Form*Z's FMZ, STL, and IGES – CCA has thousands or even tens of thousands of files it its holdings from the 25 Archaeology of the Digital projects alone.

SLIDE 24

And those are just the more common, commercially successful formats. Here we see an advertisement for Cubicomp PictureMaker, a long-obsolete 2D and 3D graphics system. In a recent accession, CCA received 504 digital files created in the 1980s in PictureMaker. A handful of these files are in standard raster formats that we have no trouble reading. Approximately 80% of the files, however, and including the original design files, are completely inaccessible except through their native software running on working Cubicomp hardware. While we are just starting to look into this problem, providing access to this material will be challenging at best. Given limited resources and time, it may even prove beyond our abilities at this point.

SLIDE 25

The PictureMaker files are a good example of the fact that we are often speaking about two related but separate things when we talk about digital preservation.

The first of these – bit preservation – is the practice of ensuring that the files remain accounted for and unchanged over time. We could have an entirely separate seminar on just the best practices for bit preservation, but in essence this involves keeping many copies of data, regularly auditing contents through the use of cryptographic hashes (also known as checksums) over time, and replacing corrupted files from uncorrupted backups as needed. Proper bit preservation practice is about removing single points of failure from our systems, so that malicious or faulty user behavior, hard disk failure, manufacturing defects, software glitches, natural disaster, and other potential risks are unable to irrevocably damage our collections.

But we often also mean something else when we say digital preservation – what I will call content preservation. This is the practice of ensuring that files remain not only recallable on demand, but also usable. Ensuring that we can continue to access the informational content of files over time is something of a different endeavor that involves anticipating and responding to file format, software, and sometimes even hardware obsolescence. Most digital preservation literature and practice suggests one of two approaches to dealing with the effects of time and technological change: migration or emulation. I will talk about these two approaches in relation to CAD files shortly but it is important to note that regardless of the strategies we choose to employ, successful digital preservation requires continued effort over the long term.

SLIDE 26

Luckily we have around 20 years of standards development and practice in the digital preservation community to draw upon for guidance. The Open Archival Information System Reference Model, above, is a familiar sight to anyone who has looked into digital preservation literature. An ISO standard born out of work by the Consultative Committee for Space Data Systems, the OAIS Reference Model outlines the core goals and responsibilities of a compliant digital preservation repository. With a little explanation, many of its concepts and diagrams are extremely helpful in providing common ground for training, benchmarking, and software development.

SLIDE 27

Other diagrams and sections in OAIS are a little less straightforward, and arguably less helpful. Nevertheless, OAIS remains a good starting point and common ground for planning our digital preservation strategies – both for bit preservation and content preservation.

SLIDE 28

The approach to content preservation that underlies the majority of digital preservation practice, research, case studies, and tools to date is file format migration. Put simply, this strategy involves two steps:

First, identifying file formats that meet our requirements for long-term preservation and access. So-called preservation formats have characteristics that make them less likely to obsolesce in the foreseeable future, such as being openly documented, widely supported, and uncompressed; while access formats should be easily usable by a wide range of potential users.

Second, migration involves making copies of at-risk files in these "preservation" and "access" formats and storing the newly created derivative files alongside the originals in our digital repositories.

This strategy often works well so long as we are able to identify a file's "significant properties"; that is, the precise informational, visual, or other content within the file that we are seeking to preserve. OAIS asks us to make these considerations based on an understanding of our "designated communities" – the primary users for whom we build and maintain our archives.

For many types of file formats, migration is a simple and effective strategy. Raster images in proprietary compressed formats can be migrated to uncompressed TIFF, which is openly documented and more resilient to bit loss than, for example, a JPEG. A word processing document such as the one above – if all we are interested in preserving is the text – can be migrated to an archival format such as PDF/A.

Of course, even for this last example, we can start to see why these questions of significant properties and designated communities matter. Whether the Word document has annotations such as comments and track changes – and whether our users need or even want access to this data over time – will have an impact on our file format policies and migration pathways.

SLIDE 29

So, thinking about migration as a strategy for content preservation in relation to architectural design, the first question we have to ask ourselves is: what are the significant properties of born-digital design records?

And the answer, of course, is that it depends on who our users or “designated communities” are. The significant properties for a 3D CAD model will be significantly different for a student who wants a visual sense of the design; a scholar who is interested in the parametric elements of a design model; and an architect or engineer who wants to use the data for a renovation project. In other words, the first question to ask is: what exactly do we want our users (and ourselves) to be able to do with these files in 5, 10, or 20 years? Beyond?

SLIDE 30

At CCA, especially for the Archaeology of the Digital project archives, curatorial interest is often directed toward the design process and the effects of technology on architect and architect on technology. Because CCA is seeking out what was new and interesting about these projects, the ability to re-create in a way the design process is crucial for our designated community.

Take, for instance, this example from the Testa & Weiser records. This screengrab, created by the CCA curatorial team, demonstrates the application of a MEL script developed Peter Testa and Devyn Weiser’s Emergent Design Group at MIT to a three-dimensional model in Maya.

In Testa & Weiser’s Carbon Tower project, this Weaver script was used to algorithmically weave ramps along the exterior of the tower. In the absence of a traditional central core, these ramps would provide one of the primary means of circulation between floors. The weaving of these ramps – and the technology that enabled it – are crucial design elements of the project, and thus highly “significant”. Yet this knowledge doesn’t necessarily make apparent a migration-based digital preservation strategy for this material, because we are dealing with a script written in a language that is specific to one software platform. We can illustrate the basics through video clips like the one we just watched, but if we want to allow researchers to more fully engage with the material themselves, there doesn't seem to be much alternative to giving them access to the script and a version of Maya within which the script will execute.

SLIDE 31

Examples such as the Weaver script aside, when we are able to identify the significant properties of digital design files, what file formats are generally available as potential preservation formats?

The two most promising "heavyweight" or fully-featured formats, cited by Alex Ball and many others as potential options – are the Standard for the Exchange of Product Model Data (or STEP) for CAD files and the Industry Foundation Class (or IFC) for BIM.

STEP describes not one file format, but a family of ISO specifications and attendant file formats that have been in development since 1984, including AP 203 (sometimes referred to as a STEP physical file) and AP 242, which expands on AP 203 with additional functionality, including visualization through approximate geometry and 3D product and manufacturing information.

While STEP -- a vendor-neutral, openly documented family of file formats – is an attractive choice as a preservation format, it does have its drawbacks. First, it is a complicated standard, consisting of multiple file formats and at one point holding the record for the longest ISO standard on the books in terms of page count and technical appendices. In addition, some data loss is inevitable in the migration to STEP, especially parametric data. Finally, not all vendors have introduced STEP import and export functionalities into their products, meaning that at this point, STEP might only be an option for a certain range of formats.

As John Gelder pointed out yesterday afternoon, IFC is an extremely promising format for BIM, with the potential to be a truly interoperable archival file format for the projects of today and tomorrow. This is certainly worth keeping an eye on, but does not help to solve the problem of what to do with data from the last 30 or so years.

Other formats sometimes floated as ideas for preservation formats – IGES and DWG/DXF in particular – also have their limitations. Neither IGES nor the AutoDesk family of formats, to say nothing of an even lighter format such as 3D PDF, are capable of retaining all of the elements of a CAD model created in a program like Rhino or Maya that we might determine as being significant.

This is likely what led the FACADE project team at Harvard to recommend not any one of these formats, but all of the above. If we don't know that any of these formats is individually sufficient for our users, perhaps we can create a wide number of derivatives, in the hopes that at least one of them will meet a user's research needs at some point in the future.

SLIDE 32

Another consideration we must keep in mind with migration of CAD files is validation of our work. As the two quotes above – from Alex Ball's report in 2013 and a much more recent project preserving archaeological CAD data at the University of York – help to illustrate, data loss when moving CAD models between software platforms and file formats is both common and difficult to recognize.

Ball suggests that this problem can be solved through recording and checking what he calls validation properties. For instance, to move File A from Format 1 to Format 2, first we open File A in its native software environment and record information such as point clouds and the volumes of solids in the model. Then we export the file into Format 2, open our new file and compare notes to ensure that the data we are interested in hasn't changed.

Two things seem apparent to me here. One is that these methods really only allow us to audit the consistency of certain types of information – i.e. the geometry of a model. But say that is not a problem, because we've determined that our designated community only needs that information. It is also true that this process is very labor-intensive, and is perhaps only feasible in repositories that are receiving large number of files to the degree that the process can be scripted or otherwise automated.

In the FACADE project, the process of creating preservation derivatives involved a "CAD specialist" who generated STEP or IFC, IGES, and 3D PDF derivative files and detailed metadata for each individual CAD model, as well as an "architecture specialist" who performed quality assurance work to assure the validity of the files and their metadata. With collections of limited scope – say a few hundred files – for which we are able to acquire dedicated project funding and personnel, this is a wonderful approach. But it simply doesn't scale to the reality of digital archives like those being collected by CCA, which may contain thousands of CAD models, interspersed throughout tens or even hundreds of thousands of other files.

SLIDE 33

Which brings us to the second approach: software preservation. Instead of attempting to create a "preservation version" of every file that comes into our archives, we could instead keep the file bitstreams exactly as they are and focus on maintaining long-term access to the software that created and can authentically render them. This entails a few things:

- Collecting contemporary and legacy software
- Dealing with licensing and End User License Agreements for the software we collect and their dependencies, such operating systems and drivers
- Implementing technological frameworks to provide access to the software on request.

This approach is not new, but it is perhaps newly viable, especially from a technological perspective. I'd encourage all of you here to read David Rosenthal's report for the Mellon foundation last year on the current state of emulation and virtualization as preservation strategies. His conclusion is guardedly optimistic: in short, the technology to provide access to software long after its original software and hardware environments are commercially obsolete already exists and is improving all the time. The core challenges lie in making the continued development and maintenance of this software sustainable, and in addressing the aforementioned legal challenges surrounding licensing.

SLIDE 34

Here we have a few examples that I hope demonstrate that software preservation and emulation is not, or no longer, a purely academic discussion.

In the upper-left and lower-right hand corners are the technical diagrams for two emulation-as-a-service platforms currently being developed and deployed: Germany's bwFLA project and Carnegie Mellon's Olive; respectively. Both platforms have their code online and can be deployed as-is in the context of a collecting institution. Both are also are discussed in detail in Rosenthal's Mellon report.

In the upper-right, we have oldweb.today, a fantastic web resource worth exploring in a little more depth. Oldweb.today combines web archiving with software preservation, allowing users to view websites harvested over the previous two decades on a range of new and old browsers. The older browsers – for instance, Netscape Navigator for Macintosh – run in emulators that are deployed seamlessly in the browser, giving the end user an authentic experience without requiring from them very much in the way of technical expertise.

And finally, in the lower-left, perhaps the most banal example of virtualization: my work laptop. Like many other people around the world, I run Windows 7 in a VMWare virtual machine on my Macbook Pro every day, an act not so dissimilar from what I'm suggesting here as one bright potential future for CAD/BIM preservation and access.

SLIDE 35

Acknowledging that there are still legal and technological issues to be worked out around software preservation, where does that leave us? Well, at present, we can do a few things:

- Collect as much contemporary software as possible, with the aim of being able to provide access to as many of the files in our collections as possible.
- In parallel with the above, establish contacts with software vendors. It has been the experience of CCA and other institutions with whom I've spoken that vendors can be friendly and supportive of collecting institutions' efforts, especially when we are able to convince them that our work is interesting and good for their brands. Ideally, we can

also take this a step forward, dealing with software companies as a community rather than as individual institutions, through projects like UNESCO's PERSIST and the Software Preservation Network.

- Finally, like good archivists, we can document this acquisition, our conversations with vendors, and the terms of licensing agreements, so that if we are in the near future technologically able to implement emulation-as-a-service platforms for preservation and access, we know exactly which software we are clearly within our rights to use.

SLIDE 36

So far I've talked mostly in the abstract. Now I'm going to switch gears and give you an overview of what the act of preserving, cataloguing, and making accessible our digital archives actually looks like at CCA, at least at this point. Keep in mind that this entire process happens within a larger context of curatorial activity, exhibition, and publication at CCA – but that today, given the audience and the focus of this session – I'm going to focus on the workflow primarily as it relates to CCA's archivists and cataloguers.

SLIDE 37

CCA's workflows for born-digital archives have changed quite a few times since 2012. In many ways, the process of developing workflows has been an experimental one, with input from a number of CCA departments. Here you can see just a few examples of how ideas about how to work with digital archives were codified and changed over time.

SLIDE 38

At present, our workflow generally consists of these 9 steps. As much as possible, we try to align our work with recognized best practices and rely on community-backed, open source software tools, although as we'll see, these occasionally need to be supplemented by relatively simple Python or Bash scripts to conduct analysis or to make sure that tasks are conducted in routine, predictable ways.

SLIDE 39

Everything begins at CCA with a general curatorial interest; in this case, in a project, architect, or firm. Unlike many other institutions, where an exhibition may be curated from a selection of what already exists in the collection; at CCA, it is typically the exhibition and publication processes that drive collection development.

Once this interest is established, there is some preliminary appraisal that happens. This can mean selecting only a number of projects or a specific type of material. In other cases, especially when digital files are on physical media that the donors themselves can no longer read, we might ask the architects to send everything they have and then conduct any more detailed appraisal once we have the material at CCA.

At this stage, a short data transfer guide is also sent to donors. This document gives donors several options for how to send their files to CCA without altering file contents or metadata such as datestamps – encouraging them to use tools such as 7zip, robocopy, or rsync.

It is also at this stage that we will provide donors with the Submission of Digital Files Information Sheet; a document more commonly referred to at CCA as the Questionnaire.

SLIDE 40

The Questionnaire is a form-fillable PDF document that asks donors to provide any information they might have to aid in our understanding of the files being donated, their provenance, and the context of their creation and use. More specifically, it asks questions about:

- The history of the organization or firm
- Staffing and roles relevant to the material donated
- The projects to which digital files relate
- The files themselves
- The computing environment the files were created in
- Any software used
- The design methodology process

SLIDE 41

Here we see the final section – which asks donors to identify the software used in the conceptual, design development, and construction phases of a project, as well as any plug-ins that may have been used.

We've found that donors are typically more than willing to provide this type of information – or at least as much of it as they can precisely remember – and that this information is invaluable for CCA's research, cataloguing, and preservation activities.

SLIDE 42

Once a deed of gift is signed and the physical and born-digital components of an archive are sent to CCA, we enter the Accession phase of the workflow. This is managed in large part by the CCA Registrar, with whom we have been refining our policies for how to handle digital materials on arrival at CCA. It is at this stage that any physical digital storage media – such as CDs, hard drives, thumb drives, floppy disks, and Zip disks – are separated and marked for further stabilization work.

SLIDE 43

In the Capture stage, data is copied from physical media, and, together with any files sent to CCA via network transfer services, is ingested into the Dark Archive via Archivematica, an open-source digital preservation ingest and repository management tool.

Data from original physical storage media is captured using digital forensics tools and techniques borrowed from the law enforcement community and refined for use in collecting institutions by groups like the Bitcurator Consortium. In many ways, the aims of law enforcement and collecting institutions are similar when it comes to capturing digital data – for both communities, the value of data being considered is dependent on its demonstrable authenticity, either to pass standards of admission in a courtroom or to serve as primary documents for scholarly research, renovations, and other activities of our users.

SLIDE 44

Most original media are captured in-house on a dedicated PC running Bitcurator -- a flavor of the Ubuntu (Linux) operating system preconfigured with software and scripts to aid collecting institutions in capturing and analyzing data – in combination with a small collection of common legacy media readers, such as 3.5" and 5.25" floppy drives, CD/DVD drives, flash media readers, and Iomega Zip and Jaz drives. Drives and devices are connected to the computer through hardware write blockers at all times, which ensure that their data cannot be accidentally altered by the capture workstation, and then they are disk imaged. The resulting disk image files – each an exact software copy of data as it existed on a particular physical medium – are checksummed and transferred permanently to the Dark Archive, which ensures that CCA always has a copy of digital data received in an accession exactly as it arrived at CCA.

SLIDE 45

Once an archive is assigned to be processed, we then undertake a phase of triage, reporting, and research. The archive is interrogated primarily through three different methods:

1. Through research conducted by CCA's curatorial teams and activities, to give context for the work and its significance
2. Through technical metadata extraction and file characterization, to give us a detailed understanding of the files on a micro- and macro-scale, and to identify potential preservation issues, such as mal-formed or unidentified files
3. Through Skype walkthroughs with the architects and/or their collaborators, wherein we share our screen and go through the digital archive with them, asking questions about how the files were created and organized, their role in the design process, and other such information.

These three methods balance and supplement each other, and in ideal circumstances give us a rather holistic work of the material before processing begins.

SLIDE 46

Here we have an example CSV output from Richard Lehane's Siegfried, one of the file characterization tools employed at CCA. Siegfried works by comparing the actual code of files against PRONOM, a file format registry database maintained by the National Archives in the

UK, to precisely identify file formats, versions, and related information. This method of characterization is far more accurate than relying on file extensions, which can be arbitrary and misleading, particularly for archives that date from before modern practices of file extension use become standardized.

SLIDE 47

Using a program called Brunnhilde, which takes this Siegfried CSV output and loads it into a SQLite database, we can then query aggregate reports out of the data, to identify trends and get a high-level understanding of the material. For example, here we see sample reports on file formats and last modified dates in the Testa & Weiser records, each sorted by count. Brunnhilde can also be used to identify duplicate files or files which raised warnings or errors while being processed by Siegfried, and can be extended through the addition of SQL queries.

SLIDE 48

All of this research is then used in pre-processing planning to develop a processing plan for the archive. These meetings include members from multiple teams at CCA, to help align processing with other activities happening at the institution.

SLIDE 49

Then the work of processing – that is, arranging and describing – begins. There are a number of factors that can influence the arrangement and description of a digital archive, but broadly we try to follow a few guiding principles.

First, “let the bits describe themselves.” Born-digital records are unique in that they already contain much of their own description, and that we can automate the process of extracting and using this metadata. Things like date ranges, notes on file formats, and extent statements can – and should – be auto-generated and only verified, not manually checked or typed out, by humans. This allows processors to focus their energy and time on uniquely human talents, such as contextualization, interpretation, and analysis, that add value to our descriptions and guide researchers toward records that might meet their informational needs.

Second, minimal rearrangement. On the one hand, this is simply the application of a time-honored archival tradition of *respect des fonds* and original order to a new medium. On the other hand, it is something of a practical necessity. No matter how much documentation we gather and no matter how thorough our pre-processing investigations, in any accession of a certain size there are likely to be linked files and other dependencies that we are simply unaware of. Rearranging material within folders can result in us breaking these dependencies without even being aware of what we are doing, and so should be reserved only for when there is clearly no original order, or when the existing original order significantly impedes access and use. In practical terms at CCA, minimal rearrangement means that we often assign top-level directories to appropriate Series and Project subseries, but do not alter the sub-directories and files underneath.

Third, co-arrangement of born-digital alongside other formats. While a separate "Digital" series is sometimes necessary, and in some other cases maybe even desirable, we proceed from the assumption that if a researcher goes to the series for X project, she should find all of the records in the archive related to that project – the paper alongside the digital alongside the models – rather than having to look in two separate places in the finding aid hierarchy.

SLIDE 50

Here we see a simple example of letting the bits describe themselves – a CSV output of a simple Python script, which walks down a directory structure and reports back information on each subdirectory it finds along the way, formatted so that it can be easily transformed into a data entry spreadsheet for a collections management database.

SLIDE 51

And an example of what our description looks like in CCA's current finding aid interface. This does not include some of the data being recorded in TMS – for example, you won't see the extent statement in the Excel spreadsheet on the last slide – but that will change in the coming future, as CCA replaces its existing tools for access to its research collection. You'll note that the file-level description can apply to materials of varying sizes, depending on the original folder structure of the material.

SLIDE 52

In a somewhat parallel process to Arrangement & Description, the final processed file-level groups are ingested into Archivematica, where they receive full preservation treatment. The Submission Information Packages ingested correspond exactly to the file-level groups described in the archive's finding aid. Archivematica then creates preservation and access copies of the files according to CCA's local file format policy, and stores these copies securely in the Dark Archive.

SLIDE 53

CCA is currently in the process of moving into more of a production environment with Archivematica and expanding its use to more of the processing staff. We are also looking into the possibility of leveraging the detailed technical metadata Archivematica extracts from files to auto-generate item-level records in our finding aids for each digital file, with the ultimate goal of having the actual files themselves viewable or downloadable alongside their fully automated, detailed descriptions.

SLIDE 54

Which brings us to the next step of the workflow: access. Once born-digital records are described in finding aids, they can be requested by members of the public for consultation in our Study Room in Montreal. An on-demand process at the moment, this involves setting the researcher up at one of CCA's four CAD workstations (**SLIDE 55**), which are loaded with a wide range of CAD and BIM software. The Study Room workstation is locked down with no internet access and blocked ports, so that researchers can see and interact with files but are unable to copy them for off-site use.

As you know if you saw my presentation yesterday, CCA also considers our exhibitions and publications to be a form of access to the born-digital archives (**SLIDE 56**). While this is certainly different than exploring the archives yourself, the shows and e-publications offer interesting insights into the material and can bring the material to a much wider audience than those who are able to physically travel to our Study Room in Montreal. Here, for instance, is a screengrab of the recently published Testa & Weiser Carbon Tower ibook, which includes a number of native digital files and images which can be viewed in galleries and lightboxes alongside the text of an interview between the firm principles and Greg Lynn.

SLIDE 57

And of course, because digital preservation is a sustained and long-term activity, there are ongoing tasks of monitoring and repair. In addition to general maintenance of the storage infrastructure, this involves checking periodically to make sure that our holdings are complete and unchanged, and responding appropriately when this is not the case. In CCA's case, these fixity checks happen through Archivematica, which first ensures that all of the Archival Information Packages in its index are accounted for in storage and then validates checksums for each package. In case any file corruption is detected, the AIP can be restored from one of several backups managed by IT, including a last-resort tape backup stored over 70km from our building in Montreal.

SLIDES 58-59

I'd like to conclude here today on a note that I find encouraging, and I hope you do too.

Although long-term preservation and access of born-digital architectural records like CAD and BIM files is a big and multi-faceted challenge, there are a lot of people out there with an interest in this work. These people come from a number of backgrounds – architecture; computer science; library, archival, and information studies; gaming; film; engineering; conservation; and more. I know from my own involvement with the Society of American Archivists' CAD/BIM Taskforce – a group whose membership has quickly grown beyond the borders of the US – and conversations with folks working on projects such as the Software Preservation Network, that CAD/BIM preservation is on the minds of many. And of course, the fact that we are all at this workshop today is further testament to this fact.

All of which is to say that, although the challenge ahead is significant, its solution will come via projects and conversations like these, engagement with those in allied professions and pursuits, and sustained effort. And the larger and more communicative the group working toward the solution is, the more success we are all likely to find. Thank you.