Legacy Data Conversion

Background

Legacy Data Definition

1. Information stored in an old or obsolete format or computer system that is therefore difficult to access or process

Problem Statement

Many of the UK's built assets were constructed prior to the digital age. The data sets behind the construction and operation of these assets exist as hard copy, 2D data or on obsolete IT systems that aren't compatible with modern BIM/PLM/AM systems.

This large quantity of legacy data potentially has a significant value to construction companies and asset owners but there is little consensus within the industry on the process of realising the value of that data in an organised and cost effective manner.

This poster explores the process of using Legacy Data by construction firms and the problems they encounter when doing so. The process has been split into 6 steps, each of which forms a research strand. The strand includes an overview of the problem, industry best practice, a breakdown of the technology available to assist the process step and corresponding pros and cons to that technology.

Finally a future state of the use of legacy data is depicted together with a development strategy to reach that point.

69%
Typical amount of unstructured data that has no business value
[Compliance, Governance and Oversight Council (CGOC) survey, 2012]

£500m/year
Cost estimate to UK government of re-creating existing data
[Better Information for Better Government 2017-01-18]

"Increasingly legacy data is used for analysing trends and using the value of data to provide predictive interventions (or to sell those insights)"
Matt Blackwell / Costain / Digital Operations Director

Use of Legacy Data Case Studies

CASE STUDY: Network Rail + Bridgeway Consulting / Liverpool Street Station BIM Model/July 2015

Network Rail commissioned Bridgeway Consulting to create a 3D digital model of Liverpool Street Station. Bridgeway were provided access to Network Rail's archive, a store of 29k+ drawings and documents related to the station design.

A desktop study of this legacy data concluded that it was not cost effective to base the model on the archive material as the process of validating the data was too complex and expensive. Instead Bridgeway created a BIM Model of the station from scratch using point cloud data generated from 600 laser scans. Data-richness was increased by supplementing the laser scan results with Ground Penetrating Radar and Mechanical, Electrical and Plumbing (MEP) survey data also collected by Bridgeway.


Figure 1: Final BIM Model / Laser scan dot cloud


Figure 2: Liverpool Street Station [Google Maps]

Key Benefits of Using Legacy Data:

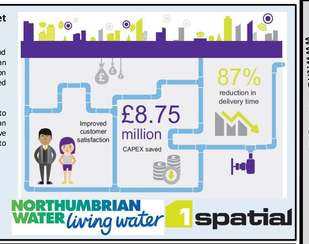
- None!

[Simon Hatch / Survey Engineer / Bridgeway Consulting August 2018]

CASE STUDY: 1Spatial and Northumbrian Water Digital Asset Management and Mapping

In 2011 the UK government transferred ownership of all private drains and sewers in England to local water companies and as a result Northumbrian Water inherited an estimated 13,500km of wastewater pipes. Information on this infrastructure was limited and to make up the shortfall NWL approached 1Spatial to fill the gaps.

A traditional surveying approach was calculated to take around 15years to complete, meaning a significant delay before NWL could plan and execute an operation and maintenance schedule for the new system. An alternative solution was required. 1Spatial's approach was to use available information to create a set of data rules to infer what the missing information should be.



Key Benefits of Using Legacy Data:

- £8.75 million saved on survey costs for transferred network
- Mapping completed in 1/8 of the time originally planned

[https://1spatial.com/news/2017/03/1spatial-enables-northumbrian-water-save-8-75m-digital-asset-management-mapping/]

i3P Member Survey Results

The survey was answered by 8 individuals from 6 businesses, evenly split between Client, Contractor and Consultancy. The survey demonstrates a diversity of practices with legacy data. There is a mix of older and more recent data formats identified in responses and, it is reasonable to conclude that as technology advances at ever-increasing rates, the industry must innovate to be able to successfully manage legacy data gathered from a breadth of disciplines through the lifecycle of an asset or facility.

mtc Manufacturing Technology Centre	BIM models	O & M system (CAFMS)	CAD model / digital drawing	Searchable database	Backup	Manual / information pack	Data validated Y/N
Amey							
Anglian Water							
Atkiss							
Atkins Global							
Balfour Beatty							
Bam Nuttall							
Skanska							
Bentley							
CIOM							
Costain							
Cosowel							
Environment Agency							
Farnwood							
Heathrow							
Highways England							
HS2							
IGP							
Kier							
Laing O'Rourke							
Mace							
Megan Sindall							
Network Rail							
Network Rail							
Osmosis							
Taylor Woodrow							
Tideway							
Vinci							

Legend
- Desired data outputs
- Not required
- No response

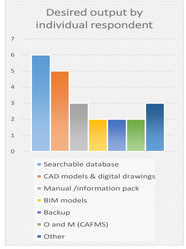
The survey was answered by 8 individuals from 6 businesses, evenly split between Client, Contractor and Consultancy. The survey demonstrates a diversity of practices with legacy data. There is a mix of older and more recent data formats identified in responses and, it is reasonable to conclude that as technology advances at ever-increasing rates, the industry must innovate to be able to successfully manage legacy data gathered from a breadth of disciplines through the lifecycle of an asset or facility.

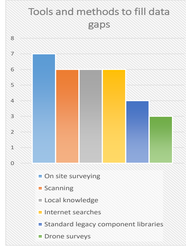
Outputs and Storage
The table shows the desired outputs from legacy data. Respondents commonly referred to a diversity of data i.e. "variety of sources", "anything and everything" and the cavernous capacity required for single formats, "…300TB of email on storage servers".

Uses and Ownership
The uses of legacy data are diverse and varied, ranging from describing existing assets, producing an effective design, planning, comparison of cost estimate v outturn, performance information, safety and quality, trend analysis, predictive interventions and warranty management. Responsibility for legacy data is typically held by the IT function, whereas ownership resides with clients and sometimes project teams.

Validation of data
The majority of businesses are using some form of data validation depending on project requirements and/or data format, although most comment that processes are manual and that data validation is not universal. Respondents indicate there is a need for a standardised, common approach and cite the proliferation of too many tools and methods as unhelpful. The volume of data and its content is unquantified/unknown and, therefore, to quantify the cost and time to ascertain value, is a barrier to progress.

Opportunities
Respondents identified a breadth of potential innovations and improvements, particularly in the design and operation of built assets, ranging from management of operating costs, feedback on energy consumption and maintenance needs and hence better informed designs. Other opportunities include identification of safety hazards, predictive analysis, monitoring internal performance and enhanced processing of geospatial survey output.





Process Step / Problem Scale / Best Practice Industry / Output OR Technology / Findings

Scoring Criteria

5	Excellent / YES
4	Above average
3	Average
2	Below average
1	Poor / NO

STRAND 1 — Legacy Data

£££ RETURN ON INVESTMENT £££

COST OF ACQUIRING DATA FROM SCRATCH

COST OF SORTING & VALIDATING LEGACY DATA

ROI 1
Validating and converting unstructured data to an up-to-date format can be expensive. In order to confirm that using legacy data is providing value for money a project manager may wish to introduce gate reviews to the legacy data process where a high level comparison of legacy data conversion costs and the cost of acquiring new data can be compared and a decision made on how best to proceed.

ROI 2
Another factor to consider is the amount of value available in legacy data and when to extract that. Taking a long term, program view, the asset owner may wish the complete legacy data set to be validated giving them a complete picture of the available data that may be used across a whole program of works. Conversely, taking a short term, project view, a developer may cherry pick only the project-specific data that they require. The latter is the cheaper and quicker of the two options but if future projects do materialise the cost of acquiring data for those may exceed savings made here.

Inspection for Construction / Wearables for Health and Safety / Robotics and automation / Standardised component sets / Fixed Asset Technologies / PLM / AR, VR, MR / i3P Discovery Projects

1. What is the output of using legacy data?
Dependent on company and perception of what will be useful to them in the future.
Once the output is specified, the value of the legacy data gathered and the additional cost to complete the output requires determining - is a certain amount of legacy data required to reap benefits (i.e. cheaper than starting from scratch?, Go/No Go)
Currently this is a manual process with the Go/No go decision being subjective.

2. Legacy data capture
The process of understanding what the available sources of data are and collecting data relevant to the project.
Typically this involves interrogating multiple legacy data systems. (electronic and paper-based) - a laborious and time consuming task

3. Validation
The process of determining whether the legacy data a company has available is an accurate enough reflection of the asset in its current state:
What is the cost of validating the legacy data you do have?
What is the cost of commissioning further data acquisition activities to fill the voids in the legacy data set?
Can these costs be spread over the project/lifetime of the asset to justify the validation exercise or would it be less costly to begin from scratch

4. Gap Analysis and Gap Filling
Method of identifying the gaps between known data and unavailable data and putting in a plan to bridge the gap with data from alternative sources;
Determine information you have against information needed for output and calculate costs to bridge the gap.
Method of identifying gap is currently manual and accuracy is dependent on user interpretation.

5. Data Conversion/ Migration
How can legacy data be converted or migrated into modern and useful formats?
Conversion from paper data to digital data is manually intensive, expensive and prone to human error. Output can only be as good as source data.

6. Output Validation
Method of ensuring that the physical asset matches the virtual asset.
How do you validate whether output is correct or not and how are changes recorded and managed?
Potentially need multiple iterations (in the form of subsequent projects over the lifetime of the asset) to obtain full validated model.

Pros / Cons

Method	Pros	Cons
CAD Model	- Relatively small amount of data; there is a higher probability that sufficient legacy data is available to produce a single CAD model then full BIM model	- Requires a legacy data to generate CAD models of existing assets
BIM Model	- Offers a time saving over surveying for all-new data required to produce BIM model provided legacy data is complete and can be quickly and easily validated	- The use of a BIM model requires a large amount of high quality data and the significant costs linked to validating large datasets cannot be justified when cheaper and simpler alternatives for collecting new data such as laser scanning are available. Instead, legacy data is best used for metadata, asset maintenance or user documentation applications.
Searchable Database	- Incorporating legacy data formats of data into a single model likely to be difficult = specialist skill set - Voids in legacy data = unlikely to have the richness of data required to generate BIM without significant amount of rework / gap filling	- The use of legacy data to create a searchable database e.g an Operation and Maintenance (O&M) database for an existing building – there is value in legacy asset management data that can be packaged and sold on to populate CAFM systems and produce more accurate preventative maintenance schedules
User Manual	- Logical first step in digesting a large legacy data set. - Allows relevant information to then be searched for when specific to the task in hand - Potential for database to accept a variety of file formats = less work required to convert data to a suitable format	- Finding a specific piece of data required to generate a model could be difficult/labour intensive & a gap fill exercise may be required - Complex transition from multiple legacy systems to a new, single system - Substantial amount of information to convert to electronic (including search terms). Would require a significant amount of labour to do manually or programming to do automatically
Auto Data Harvesting	- Historical data can reveals flaws in a product that can be rectified with a preventative maintenance regime documented in the user manual - Simple and relatively low cost to produce compared to the other outputs	- The use of a legacy data to create a user manual for an asset's systems e.g hot water system operating manual. (Trend toward paperless with electronic copies maintained online)
Manual IT searching	- Fast & easy to locate required information - Easy to search on any piece of data - Improved search experience - Potential ability to capture metadata increased	- Limited information specific to asset, ignores integrated systems(i.e. user manual for boiler would not consider radiators and other ancillaries)
Local Knowledge	- Some Manual Search processes are easily accessible e.g keyword search via Windows Explorer - Relatively low-skilled work	- A process to automatically identify and classify information - it takes content, in any type of format and converts it into readable and analysable information. There is a large number of data mining software and applications available for legacy use.
Artificial Intelligence (AI)	- Immediately available and tailored to the enquiry - Very low cost	- Heavy database architecturing - Definition of data quality standards - Model/system complexity - Significant investment in skills and technology required for most applications
Speak to people (local knowledge)	- Faster & easier to locate required information - Increased search satisfaction - Potential ability to capture metadata increased - Ability to sort and analyse technical data	- Laborious + time consuming + expensive - Legacy record IT systems could require legacy skills to manipulate that may be difficult to acquire - Prone to human error - Very limited durability - Specialised knowledge is hard to share
Site Surveys	- Low cost - Simple - Expansive information = likely to acquire better picture of the situation when speaking to an individual. - Less prone to misinterpretation	- Significant investment required in skills and technology for most applications - Limited in application at present (bespoke solutions) - Generally requires significant data sets to 'train' AI tools to recognise and structure information
Pay-to-Access Databases	- Labour intensive - Risk that poor quality data may be validated - only the memory of those interviewed (no hard data) - Data can be biased by an individual's experience - Retirement/unavailability of knowledge holders	- High quality data available provided a robust change control process has been adhered to throughout the duration of the project - Change control documentation may not be in a digital format therefore could be time consuming to find the information required - If change control process has not been strictly adhered to, the quality of data obtained from the change control documentation may be poor - Potentially dealing with commercially/security sensitive data that may be difficult or time consuming to manage
Public Internet Sources	- Potential to significantly reduce quantity of validation and gap filling work required = time + cost saving - Experience using sampling in the manufacturing industry that could be tapped in to	- Anything other than a full sample risks poor quality legacy data slipping the net and being incorporated into future work
Manual conversion	- Semi-automated technology = time-efficient - Quick gap assessment can be obtained - High TRL in education context = proven technology and low cost to implement	- Labour intensive at the front end when defining search criteria - Small pool of people with the right skills to reconfigure software to this application
Optical Character Recognition (OCR)	- Large store of information that is easily accessible - Low skilled task - Low cost - Relatively accurate information - High Speed = information available instantly	- Limited site specific data available - Documents are likely to be in scanned forms rather than in digital format
ReCAPTCHA	- Large store of information that is easily accessible - Low skilled task - Low cost - Latest information is easily accessible - High probability that information is already in digital form	- Data may be too generic - Conflicting information available from a range of un-verified sources therefore could be labour intensive to interpret and validate data
Digital Conversion Platforms	- Common method to convert paper drawing to digital format (CAD drawing). - Many companies already established to provide this service to industry. - Customer can specify end format required.	- Labour and time intensive. - Only as good as the source drawing. - Requires interpretation which may be prone to errors. - Only works if paper drawing or record is available.
Nuance	- High TRL - this technology has been widely in used for quite sometime (i.e. passport and Automatic Number Plate Recognition ANPR). - Well understood. - Efficient. - Allows documents to be searchable after scanning.	- The software requires formatting to ensure that it picks up variances of how the information can be presented. - Works best with good quality typed documents; handwritten documents are more difficult to read. - Information is only as good as the source material. - OCR has difficulty differentiating between characters (i.e. number zero and capital 'O') - Requires manual clean-up which can be prone to errors.
Theorem Solutions	- High TRL – has been in used for sometime. - Semi-automated as it only requires human intervention when the OCR system cannot read the text - Self verified and high degree of confidence	- Only works with text and simple images - Relies on collecting data from multiple users - The software requires adapting if used for to digitise 2D paper drawings
	- Several companies and software exists to convert old format digital drawing to new (for example Catia V4, CADvertor). - Conversion to 2D and 3D possible.	- Only converts from old format drawings to new. - Full capabilities and limitation of these software's still not yet understood.

Strand 1 – Define Outputs
The Research Suggests...
There are a wide range of applications for the use of legacy data in the construction industry. As the Liverpool Street station case study shows, using legacy data for generating CAD or BIM is of limited value - the complexity of these models requires a large amount of high quality data and the significant costs linked to validating large datasets cannot be justified when cheaper and simpler alternatives for collecting new data such as laser scanning are available. Instead, legacy data is best used for metadata, asset maintenance or user documentation applications.
Next Steps
1.1 Decision Flow Chart - i3P should create a framework/process flow and best practise guide for decision making on legacy data investments that includes indicative costs, potential lost times, advantages/disadvantages and risks & responsibilities. A cost-benefit tool can be used by project contractors and suppliers to demonstrate to asset owners the benefits of proactive data collection and validation across the whole asset life. Furthermore the tool should be used to determine how the cost of legacy data exercises be split between an asset owner, operator and the contractor.

Strand 2 – Data Capture
The Research Suggests...
Legacy data can be obtained from either an electronic source or from a physical hardcopy of the original document. Searching for and extracting data from a large volume of hardcopy documents will obviously be costly and laborious and there is a limited number of tools available to do this. Alternatively there is a mature set of data capture tools available for the study of electronic data. These require a specific skill set to operate and there is a limited pool of people available inside and outside of the construction industry with these. Skills and job roles need to be clearly defined by the industry to allow the efficient management and use of electronic legacy data.
Next Steps
2.1 i3P to commission a demonstrator project through the catapult network to show relevant, practical application of existing industry tools to the C&I industry.
2.2 Create a data analytics/big data processing job function in construction - i3P should create a high level job specification(s) describing the role and defining the skills required by an individual to support legacy data activities.
2.3 Develop a training program, including input to apprentice/grad programs, to develop the skills needed in future.

Strand 3 – Data Validation
The Research Suggests...
How to validate legacy data and how to do so cost-effectively are questions that have appeared repeatedly throughout this study. A typical situation is a project where there has been a large amount of change requests that have been poorly managed. Here there is likely to be multiple versions of the truth, where the challenge becomes ascertaining which aspect of each reflects reality. A sampling approach could be applied to validate the true information, for example the Northumbrian Water case study where sufficient confidence was achieved in a new model generated from legacy data by conducting a limited amount of sampling which resulted in significant cost and time savings. The development of site inspection technology that may be used to validate data can be found on the i3P Inspection for Construction Poster.
Next Steps:
3.1 Best Practise Sampling Technique - i3P should engage with the Catapult to benchmark best practice from manufacturing sampling methods and develop tools to carry this across to the validation of legacy data in the construction industry. Info this could be an easy tech transfer project, e.g. AQL for construction.
3.2 Digital Catapult Engagement - i3P to engage the Catapult and investigate automated comparison of multiple versions of data to ascertain the most likely true version or amalgamation of inputs to create the true version.
3.3 i3P to consider secondment of staff into manufacturing quality teams in order to understand the application of Statistical Process Control (SPC).

Strand 4 – Gap Analysis
The Research Suggests...
There are mature gap analysis tools available such as Envivo and Nalytics. Using these it is possible to generate visualisations showing the links between data sets and thus it is possible to perform a gap analysis on a legacy data set and the outputs that you want from it. Again, as with Strand 2, there is a limited pool of people in the industry with the IT skills required to operate these tools. This skills shortage will need to be addressed (see strand 2 next steps).

Public internet resources and pay-for-access online databases such as Google Maps and BIMStore provide good sources of information for filling gaps in data sets. If data is not available, either electronically or in hardcopy, it will be necessary to collect new data. For an overview of the technology in this field see the i3P 'Inspection for Construction and Infrastructure' Poster.
Next Steps:
4.1 Dissemination - i3P should create awareness of the gap analysis methods, tools and techniques currently implemented throughout Construction.

Strand 5 – Data Conversion
The Research Suggests...
Once a legacy data set has been validated it is necessary to convert the data into a current format useful to all project stakeholders. Hand written documents are the most difficult to convert into a suitable format however technology is constantly improving. Those documents that remain impossible to convert automatically must be converted manually.
When converting between file formats, e.g. updating electronic legacy data to a new, current file format, there are application specific solutions currently available however the ability to convert files cross-application is not widely available.
Next Steps:
5.1 Black Box Exercise - MTC propose a collaborative piece of work to explore and produce a 'black-box'- like approach to input various forms of construction data and output a file of a desired format. This project was voted for by i3p members at the MTC and we hope to develop this black box project in 2019
5.2 Test new tool with real project data provided from i3P
5.3 Pilot use of new tool on live i3P project
5.4 Disseminate to industry with a suitable training package

Strand 6 – Output Validation
This strand is covered by the 'Lifecycle Approach to Data Collection and Management' and 'Inspection for Construction and Infrastructure' i3P discovery project posters. These posters provide an overview of the best practise inspection methods that exist across all industries and tools for data management throughout the project lifecycle. The posters conclude with a development strategy for each theme.

See INSPECTION FOR CONSTRUCTION AND INFRASTRUCTURE i3P discovery project poster

See INSPECTION FOR CONSTRUCTION AND INFRASTRUCTURE i3P discovery project poster

See INSPECTION FOR CONSTRUCTION AND INFRASTRUCTURE i3P discovery project poster

See Lifecycle Approach to Data Creation and Management i3P project poster

www.the-mtc.org/_literature_152525/Lifecycle_Approach_to_Data_Creation_and_Management_i3P_discovery_project_poster

www.the-mtc.org/_literature_152524/Inspection_for_Construction_and_Infrastructure_i3P_discovery_project_poster

STRAND 2 STRAND 3 STRAND 4 STRAND 5 STRAND 6

The NHS story so far:

1992 — "What?" In 1992 the NHS set out to computerise its patient records. This involved the mass conversion of its legacy paper-based systems to an electronic format. In 2018 this programme of work still rumbles on at a cost of over £2bn…
Why Paperless?
- Provide patients with more control over their health and well-being.
- Empower carers.
- Reduce administrative burden for care professionals.
- Support the development of new medicines and treatments.

The Scale
- Established in 1948, it now employs 500,000 staff and it caters for 64.4m people in the UK and 54.3m people in England alone.
- There are over 100 foundation trusts that have historically stored patient records in paper and recently in various digital formats.
- Every foundation and hospital has its own system of storing information.
- According to reports NHS trust spends between £500,000 and £1m per year on paper, including the cost of storing it and moving it around the NHS network.

£4bn (investment in 2016)

The lessons learnt:
- *Lack of standardisation across individual hospitals and trusts: each hospital has a multitude of different departments and specialities, all with their own ways of collecting, sorting and storing data.*
- *Different sorting/naming systems between NHS organisations: the lack of standardisation across different organisations in the NHS prevented the effective share of records between computers.*
- *Lack of a single parent IT system: the failure of the National Programme (despite the years and the level of financial investment) demonstrate the difficulty of the tasks involved to achieve a paperless NHS and the scale of resources required to achieve it. It also highlights the importance of identifying the best solutions at an early stage.*
- *Financial issues: continuing strain due to the need to find 'efficient savings'.*
- *Sorting through and storage of complex data: range of handwritten historical notes and the lack of accurate, reliable technology for getting it right first time.*
- *Lack of staff engagement: the change of culture for staff that are used to doing things in very specific ways for their entire careers may be reluctant to constant sudden step changes driving down the uptake of new technology.*

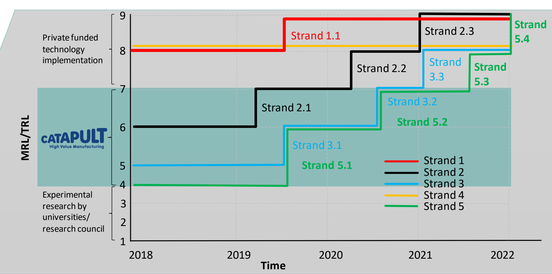
1992 — Then-health minister Virginia Bottomley announced the NHS IMBT strategy "Getting better with information". It promised to deliver a fully computerised NHS and set a series of goals for 1995.
1998 — Then-health secretary Alan Milburn published 'The Building the information core' strategy, an IT programme that would 'build on and update' the 1998 strategy. It promised fully electronic, shared records for all patients by 2005, accessible by patients and integrated across all providers by 2005.
2001 — Then-health minister John Hutton published 'The Building the information core' strategy, an IT programme that would 'build on and update' the 1998 strategy. It promised fully electronic, shared records for all patients by 2005, accessible by patients and integrated across all providers by 2005.
2002 — Milburn and other ministers then announced the 'National Programme for IT', an ambitious (and it turned out) undeliverable) project to move the NHS to a single, centrally-managed electronic care record by 2010.
2003 — Then-health minister John Reid announced that every NHS patient would have an individual electronic 'NHS care record' by 2010. Now known as the 'Summary Care Record'.
2007 — The Public Accounts Committee report, NHS IT, was critical, saying that patient records were still 'mainly on paper' and even when electronic, they were not shared between providers.
2010 — Coalition government dismantle the National Programme for IT, ending its contracts early. The project had cost over £10bn when scrapped and is criticising paid for even now.
2011 — Jeremy Hunt promised patients the right to access information about their hospital prescriptions and access to GP health records online by March 2015. Digital health records to be fully available across NHS by April 2018.
2013 — The Public Accounts Committee expressed doubts over the validity of the NHS's new paperless plans.
2013 — Jeremy Hunt announced that NHS patients would be able to access their GP records digitally in real-time by 2020.
2014 — NHS promised all patient records available digitally in real-time by 2020.
2017 — Jeremy Hunt abandons 2018 target, declaring it no longer possible.
2020

Development Strategy

ADDITIONAL PROJECT INFO:
All i3P discovery phase Z projects are match funded by the i3P and the MTC. Each discovery project is carried out over a timescale of 6-8 weeks, which makes use of input provided by the i3P within this timescale to support the projects. A review of industry best practices / technologies for each discovery project has been down selected according to criteria agreed with each discovery project sponsor.

Manufacturing Technology Centre

Staircase Development Diagram



Production Concept

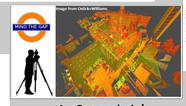
Scenario: A developer has been approached to build an extension to an existing hospital ward. The project includes refurbishing and upgrading the existing ward together with the new build element

- [x] Enough legacy data available to create output? OR
- [ ] Start from scratch?

1 – Define final output

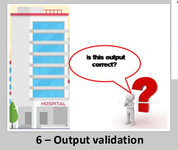

2 – Data capture

3 – Data validation



4 – Gap analysis/ Gap filling

5 – Data conversion



6 – Output validation