

Chapter 2

Document, information, data, content

How to model information?

CATHERINE LELOUP

Independent consultant, France

Ce qui se conçoit bien s'énonce clairement

Boileau. French philosopher, 18th century

What is easily understood is easily explained
[free translation by the author]

[Editorial note: 'document' in single quotes is used in this chapter to distinguish the general concept of publishable product from the more specific meaning of paper support.]

Companies often face difficulties in selecting an appropriate strategy to manage information in a way that will meet all their requirements. What are the frontiers between information, data, document and content? To what extent can electronic documents and their content be integrated with legacy systems and other traditional information systems?

The complexity of such issues probably lies in the variety of available technologies, information structure and user needs. Lack of well proven methodologies to model information as well as the weight of habit and IT history does not help. Even worse, design methods are strongly connected to technologies. A relational method is devoted to building relational databases, whereas object methods handle basic object properties and methods that are specific to the object world.

This chapter intends to formalize a possible approach to modelling information through in-the-field observations, illustrated by a practical example in the banking sector.

Clarifying concepts

Background history

There is general agreement on the definition of data as structured information. In fact, this only refers to information encoding depending on its content in order to ensure that the data is properly entered and stored in digital systems. Indeed, data is either dates, numbers of all kinds, Boolean information (yes or no), text (short or long) or binary data. This is what we have dealt with for a long time. Obviously, compared with the real world, this is a fairly restrictive view of our environment.

Then comes the world of unstructured information, for which there is no definition, only a default one - it is not structured. Indeed, it covers anything that is not data: images, electronic 'documents' produced by proprietary software, e-mail messages, graphics (assuming they were not produced using CAD/CAM software tools), sounds, and so on.

The frontier between structured and unstructured data has been tightly closed for a long time, not only from a technological viewpoint, but also as far as human skills were concerned in the computer departments of organizations. Things were quite simple 30 or 40 years ago: an automated system was complementary to other information sources such as paper documents, individual or shared knowledge. Recent terminology such as content management implies that there has been a slight opening of this frontier. However, this also is very confusing.

The *Cambridge Dictionary of American English* gives the following definitions:

Content	The subject or ideas contained in something written, said, or represented
Data	Information collected for use
Document	A paper or set of papers with written or printed information, especially of an official type
Documentation	Official papers, or written material that provides proof of something
Information	News, facts, or knowledge

Of course, such definitions were established regardless of the evolution of information technology or its possible impacts. However, it is particularly interesting to notice that these definitions mix many different basic concepts:

- the subject
- the medium, e.g. ‘paper’
- the communication channel, e.g. ‘something written, said, or represented’
- the usage: ‘collected for use’
- the source: ‘official papers . . .’
- the legitimacy, ‘proof’.

A rapid analysis of such definitions in other cultures (for example the French one, which I know the best) shows that such definitions are not shared worldwide. For instance, we call documentation a ‘subset of documents collected for a given aim and purpose’ and document ‘a physical support and the information it contains’.

Does it make sense to look for definitions? Probably not, until recent years. The web and other communication media have changed the rules. Indeed, they reveal to the external world the way internal information systems are built and how they work.

For more than five years now, the computer industry has convinced every company that it should ‘go on the web’; easy to say, but difficult to do it properly.

Content, in a web environment, is seen as a part of marketing, a means of directing users to and through a set of constantly changing links. It is a very restrictive view.

Meet our banking company

What is the business problem?

The bank wants to provide its customers and employees with consistent information about its products (a few hundred of them), at the same time, through many communication channels.

How is it done today?

Two different departments (at least) are in charge of publishing information – to the customers on the one hand (marketing department) and to employees on the other hand (organization department or

set of departments). Consistency is managed manually and based on the goodwill of each publisher.

Web and other media authors copy information such as tariffs, date of application and reference numbers from the legacy system(s) into word-processed documents and transform them (technically) before publication on the web, on paper, on WAP, and so on.

It is not only a costly method, but it is also inefficient.

How to improve the system?

The only way to improve the system is to implement a common source of information and publish it differently according to each audience's needs. Each audience receives information through different communication channels, but coming from a common source. Do not focus on web technologies only. Paper is still alive. And other media could be developed.

This example shows that it is simply not possible to redefine how information is handled each time the organization faces new challenges, whether new media, new use of information, new audience or whatever.

Therefore, we need some definitions in order to understand what we are dealing with.

Tentative definitions

Logically, definitions should address content first, information next and finally 'documents'. Data is a clear concept in the IT environment.

Content is an object with embedded rules. Contents are a set of content.

For instance, an address is composed of a street name, a postcode, a town, a city and a country, and each of these elementary components has its own rules. A paragraph is a set of sentences, each sentence being a set of ordered words separated by some special signs and beginning with a capital letter (in the English language convention). A product designation describes an individual product; each product has a different content.

The most difficult thing to define is information. News, facts, opinions, ideas, orders or knowledge could all be some sort of representation of information. However, such a definition mixes content and management rules. Indeed, a piece of news is static information. Knowledge is not.

Information is a mix of content and management rules and responsibilities.

For example, the content 'product designation' is handled under the responsibility of the marketing department, whereas the content 'product reference' is under the responsibility of the manufacturing department. The product designation may change every day with limited controls, whereas changes in the product reference are supposed to be kept to a minimum and are carried out with great care because of their possible impact on the supply chain.

A 'document' is a mix of information, publication support and communication methods

Indeed, 'documents' are made in order to be understandable by human beings. As binary data is not their mother language, information has to be processed before being published in a convenient medium.

A 'document' is not information. There are many ways to provide the same information using different 'documents'. For instance, to be informed about today's weather, you can turn your TV set on and look at the meteorological bulletin (a video scene), ask your neighbour (a sound), search on the internet (the web) or open your window (an image). At an organizational level, documents are highly redundant and probably inconsistent. It is often said that so-called electronic document management systems have failed in many cases. If using electronic document technology just consisted of transforming paper documents into electronic images, a success would have been a miracle. Indeed, depending on the role of an existing paper document - simple information support, proof, initiator of a procedure, and so on - it may become a wide range of things in a digital environment - a record in a database, an event, a message, an electronic document . . . or nothing.

Back to our banking company

The product catalogue content should assemble contents from various origins:

- a content repository (product description, legal information, marketing information, and so on), mainly composed of textual information
- data from legacy systems (tariffs, rates of all kinds).

and should refer to other publications such as procedure manuals or contract forms for the bank personnel. When publishing the catalogue, contents should be distinguished according to their accessibility to end-users (personnel of the bank or customers).

To put it another way, a final product description, whether published on the web or on paper, is the result of a process of analysing the catalogue's contents and arranging them according to publication standards tailored to their intended medium.

Modelling information: the traditional approach

Figure 2.1 illustrates the traditional approach for information management systems.



Fig. 2.1 Traditional approach of information system modelling

From the general objectives and identified constraints - technical, human, organizational, etc. - the first step in modelling an information system consists of designing future systems in terms of processes, players and requested software features; then selecting host technologies, such as web architecture or hardware platforms, and enabling the appropriate software. At this stage the information that the system should manage is not described in detail. Information models are generally limited to the description of meta-data, files and media to be handled by the systems.

This approach, although widely used, suffers from a number of drawbacks, especially when the system to be implemented deals with 'documents' - and most do. Applying our definitions we have:

- a poor description of objects to be managed
- a future and sometimes mythical system, the design of which is based on existing practice
- a lack of understanding of technologies.

A poor description of objects to be managed

In user specifications, one often finds a description of 'documents' to be handled as office automation files, HTML files and other media-specific file formats. 'Documents' are not only files. Just take a look at an office 'document'. It contains a lot of metadata (we carried out a lot of research work to find out that the only document that has almost no metadata, except a reception date, is an anonymous letter . . .) such as reference, author, title, date, subject, and so on, which are embedded in the file. The file probably includes cross-references to other documents, revision history, authentication items such as approval signatures, and so on. Of course, you can put all that in an MS Word file. But you may encounter difficulties in maintaining the consistency of versions, the validity of cross-referenced documents, just as your webmaster experienced it on your website. Because rules apply - for example version 2 can no longer be produced once version 3 has been published - the approval process depends on the document type, and so on. Some of the rules may be implicit; for example, the date implies the version.

Product description in our banking company

A product description is never unique. It often refers to information that can be applied to several products of the same family, for instance legal or fiscal product environment. It does not make sense to draft the product description from scratch each time. On the contrary, re-use of common information is essential in order to publish high-quality, consistent information, and also to control the updating processes. Significant parts of the product description should be considered as content, which should be propagated from a product family to its individual products.

In the real world, however, it is not that simple. For instance, if the law defines how a minor can subscribe to banking products in general, it is not valid to apply certain information to a given product if a minor cannot subscribe to this particular product, although other parts of the product family may be available to minors. So, propagating content should be monitored by an information system that knows which type of customers can subscribe to which products. Of course, a cut-and-paste facility could do the same, but not at the same price and not with a sufficient level of confidence in the quality of information.

The weight of existing practices

It is very difficult for users to review their current practices; nobody likes change. But what have we done with information for 40 years? We used technologies for what they could do, not for what we needed. Databases were initially implemented to compute figures, then to manage data, and are now used for a bit more than that. Office automation has enabled anyone to produce anything, without the help of any management tool except Windows Explorer, and regardless of the best practices that secretaries have established over centuries. E-mail systems are extremely useful, assuming they are correctly used, but they should not be used as electronic copying machines, with which anyone sends (receives) anything.

Personal computing and legacy systems have not always been good companions.

Corporate imperatives in our banking company

Writing the product description before the product is launched is obligatory, but parts of that description may be in legacy documents and must also be monitored in relation to what is already contained in customer contracts. Information managers must ensure that information is of high quality and complete, especially when it is automatically published on the web.

Generally speaking, existing practice regarding information management is not very good, so companies should capitalize on their know-how, not only on their current available content.

Lack of understanding of technologies

There is a magic formula in the information management field:

New technology + Old organization = Vast disaster

There are many varieties of content management technology. Any vendor will explain how easy it is to create, publish and maintain information content, but they all keep silent on their content models and embedded features. Unfortunately, these are the key to the success of a content management system. XML is not a content model: it is a grammar. It tells you how to write well formed documents, not how to design your own content model; that is clearly the responsibility of the customer, who is the only one capable of making decisions on what should be managed and how.

Qualifying information in our banking company

Everyone knows that editorial items such as tables are represented differently on the web and in paper documents. So, our XML model includes a table 'model'. But publishing the same source information to customers and internal staff requires that content is qualified either as 'public', or not; some items should be emphasized and marked (tagged) appropriately. Figures or links to other products should be managed properly in the corresponding XML model and this must be included as an option in the table model.

Modelling information: an alternative approach

Obviously, the approach should be more structured. Figure 2.2 presents the suggested approach.

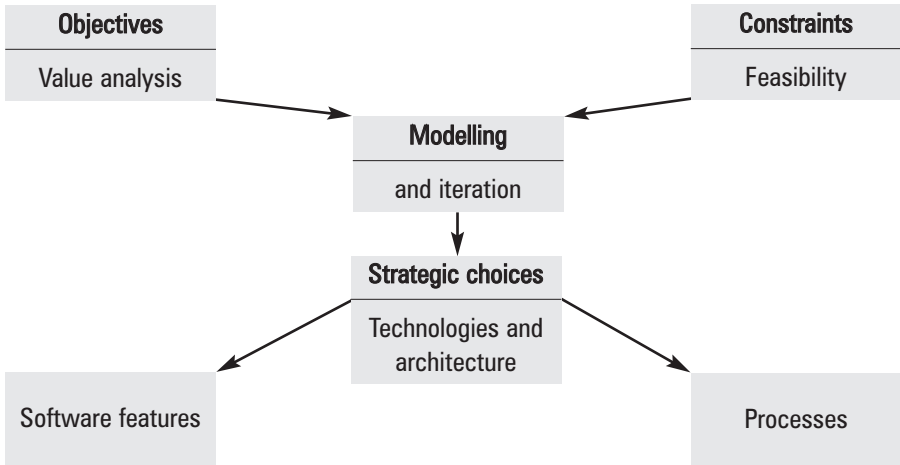


Fig. 2.2 Suggested approach to information modelling

Objectives should be defined in the light of the information needed by the company. Constraints should qualify the feasibility of the future system. Information modelling is the core issue and often requires iterations. Once information modelling has been completed, then the necessary software features and optimal processes should be defined.

Let us describe in detail each step of this approach.

Objectives: the value analysis

The question is: what is the value of this information for the company? This leads to four different questions:

- What kind of information is it?
- What is it used for?
- Who uses it?
- Who manages it?

What kind of information?

The ‘What kind of information is it?’ is the never asked question, probably because users never ask this of themselves. But it is essential. No one would manage news, facts and knowledge in the same way. From our experience, there are five basic types of information, which show very different levels of structure, management processes, volume throughputs and sharing capabilities. These are:

- *Administrative production* – Basic paperwork, from insurance contracts to administrative declarations, manuals, orders, claims, and so on. It is the field where documents are information supports and process initiators. Exchanges are essential within and outside the company. It concerns a limited but often widely dispersed population within the organization. Forms have gradually changed the working environment.
- *Reference information* – The real information assets of the organization, giving details of its knowledge and know-how. Product catalogues, methodological tools, procedures, operating manuals and quality documents are examples of reference information. This is – or should be – highly structured information, in a limited volume, used by anyone, but managed by few people. Reference information, although often updated, is a long lasting asset. Reference information could be either operating information or knowledge, but they should not be addressed in the same way.
- *Project information* – Information about a project, an activity that has a limited duration of a few weeks to a few years. Volume throughputs are huge at company level, but may be limited at project level. Few people should access all the information about a project, but many should be aware of the status of the project. Information is partially structured. A project could be anything, such as a new product launch, a computing project, a crisis project, and so on. Confidentiality is often a key issue within the organization.
- *Intelligence information* – Used widely in business information and technology surveillance. The complexity lies in the fact that from a raw magma of structured and unstructured content – web pages, professional databases, records, facts and figures, and so on – information specialists produce high value added reports. Input volumes might be very high, but intelligence reports are not very numerous and, depending on

the organization's policy, are more or less widely distributed. Knowledge databases, to some extent, may be considered as intelligence information (for instance research reports).

- *News* - By definition designed for immediate consumption. Pretty unstructured, news often feeds other information repositories, once correctly processed.

Different types of content simply cannot be handled the same way. No one would try to find something from reference information using full-text searching techniques. For example, a fireman in a chemical plant would certainly not search for the emergency protocol 2805-20, which relates to electrical fire, using a complex Boolean search like '[fire or flames or smoke] AND electric AND NOT supplies', because he would know the reference number of the document. In the real world, many people are suspicious about news, and the accuracy of information distributed on the web is often questionable.

What is the information used for?

The second question, 'What is it used for?', identifies what the key information processes are: reading and printing, re-use, sharing, publishing, distributing, relying on them to do something, and so on. There is not a one to one correspondence between use of information and information type, but this question helps to define priorities and constraints.

What is product information used for in our banking company?

A product description has many different uses:

- For the customer, it helps to understand what the product is and to compare it with competitors' offers.
- For commercial staff, it helps to sell and to point out the benefits of a product compared with the competition.

Is it compatible? Yes, but only if the company knows its competitors well. Which means that building a unique catalogue that will be published internally and for the general public does not make sense if the company does not know its competition and that knowledge is not available internally.

Who uses it?

The third question, ‘Who uses it?’, should help in identifying the actual population of users. Again, we are talking about content, not documents. And the content that is sent out to customers should also be available to the internal staff, in exactly the same way.

What was the concern in our banking company?

Technicians do not write the same content as marketing people. A comprehensive description of a product should logically help to identify commercial benefits for the customer. In the banking industry, it is a matter of money, security and return on investment.

So we had to find a way of translating a technical advantage into a commercial benefit. It consists of changing a given identified phrase by another during the publication process. XML has helped a lot. Further development may offer the possibility to automate this process as a function of the intended audience.

Who manages it?

Despite many conferences and papers, a number of organizations have not yet understood that publishing information is different from producing it. Surprisingly, fundamental rules are disregarded in connection with elec-

How to cope with this in our banking company?

One of the problems we faced was that the existing product information system was designed for internal purposes only. Unfortunately, the content was not protected against the use of ‘copy and paste’. And guess what happened? Part of the information was published on the web, by a webmaster who did not succeed in getting the information from the person responsible in marketing on time. So, he did his best . . . but it was not very useful for the customer

Company information is not an open market. Rights and duties should be clearly set, known and agreed for content production, publication and use.

tronic publishing. Producing information is the responsibility of managing accurate and updated content; publishing is the responsibility of tailoring the same content to various audiences.

This question will certainly raise a lot of other concerns, such as the legitimacy to produce or publish the information, the quality of information, its usability for a given audience, and so on.

Constraints: the feasibility issue

Naturally, human constraints - competences, skills, availability - and technical ones - integration to the IT environment - arise. They are not necessarily the most critical issues. The critical ones are:

- content quality:
 - How reliable are existing contents? Should we migrate or rewrite them?
 - How usable are they? Should we reorganize or rewrite them?
 - How accurate and up to date are they?
- organizational issues:
 - Who will manage the project?
 - Can the manager get enough resources?
 - Has the manager the necessary authority to carry out this project?
 - Who will write the content? Who has the skills?
- cost issues:
 - Who will establish the budget?
 - Would that budget be acceptable?
 - Who pays?

There are no unique answers to these questions. However, before entering the information modelling process, they should be answered. Budgets are often good regulators of the information-modelling process.

Information modelling: the key issues

The key issues are:

- satisfying objectives and constraints
- sharing the same vision among participants in the project

- explaining technologies
- validating future processes.

There is no well proven method. There are methodologies but they are for designers only, not for users, which is why it is necessarily an iterative process. Examples of questions include: Should I structure this content? Would authors be able to cope with it? If not, should we establish a clearing house or a writing house? - But that is related to processes. Could the technologies help?, and so on.

Modelling product information in our banking company

Fiscal information is particularly interesting. In France, there are far more tax rules than varieties of cheese! Basically they depend on your legal status (individual, professional, commercial company, and so on) and the source of revenue (income or real estate, for example). Logically, depending on the product family, a specific rule should apply to each legal status. But we also have more generic fiscal rules, such as death duties, which include various sources of revenue. So, in order to determine which fiscal rule should be applied to a given product, you have to know which type of customer owns the product and which type of revenue is involved, and then apply generic rules such as death duties.

These contents are recorded in a database (and then encoded) and they control the production of the product information sheet. Propagation of contents for a product family depends only on the content in the database, which is uniquely updated and applied to all the products in a product family.

Information quality is ensured by a mechanism that guarantees that a fiscal rule cannot be attached to a product if it has not previously been attached to its product family.

There is no single way to model information content. Logic helps. The company environment is far more essential. A product description in the banking industry has nothing to do with an electrical product description or a drug description. This is probably because market knowledge is essen-

tial. The only efficient way to model information for a given company is through a joint effort of information designers and market specialists.

Strategic choices

Strategic choices depend on the technologies available. One should select technologies first and then software products. Some of the considerations are discussed below.

- *Content model* It is a matter of compromise. Structured textual information requires a content model definition using XML grammar. If you want your texts to be processed, then they should be processable. XML, well formed 'documents' are processable because the schema(s) or DTDs define a logical structure of the 'document', regardless of the publication format. Authors must comply with this structure, instead of focusing on information presentation as they would naturally do with their word-processing tools. Optimistic people call it computer-aided writing, and the rest of the world calls it constraints.
- *Content storage model* A true project is an integration project, which may include native XML, XHTML, XML export-import capabilities or HTML. This means that there will be some specific software developments. It is obvious: XML is a grammar not a content model. In other words, it tells you how to build structured content, not how to specify it. Other choices may lead to the adoption of 'off the shelf' software products. HTML is still unstructured information and will never be re-usable as it does not show any semantics except title, body and metadata; it is 'flat' information. XML is a structured information content representation. Then there are cases where re-use of content is not a real issue. Standards need to be considered, also; the ad-hoc standard may be Microsoft but there might be questions about the strategic factors surrounding the use of (for example) MS XML as opposed to generic XML.
- *Workflow* Workflow is a technology that aims at automating the circulation of information and tasks among users. The key issue is to determine whether the workflow will monitor the distribution of work among users or whether each user applies rules to select other users who should contribute. In the first case, the workflow is a production workflow, in the second, it is an ad-hoc workflow. Corresponding soft-

ware tools have few elements in common except their routing and monitoring functions.

- *Publishing* Is publication automatic or manual? Is the information model powerful enough to be sure that automatic publication will provide error-free results?

What did our banking company do?

We selected XML-based formatting as a strategic choice, for various reasons. One of these was so that it would compute an automatic synthesis of our product description, re-using information from the product description database, generated automatically, just as MS Word does with a table of contents.

We did not implement workflow, because most participants involved were simply not ready to comply with such rules. We just record all steps of the validation process. We publish automatically, in any media, because this was the main objective.

Software features

These are comparatively easy to establish because however you model your content the functions to be managed are always the same. Besides the usual functions such as user management, system management, security management and day-to-day operation, there are 12 end-user functions to be specified:

- creating
- previewing
- formatting
- publishing
- distributing
- storing
- searching
- viewing
- printing
- revising
- archiving
- exporting and importing.

Once the modelling phase has been passed, this is a very simple exercise, which only requires listing what is in everybody's mind. Drafting technical specifications may require time but, once the model has been completed, it is not difficult. It is just a matter of logic and writing ability.

What did our banking company select?

Given that the production of content is monitored by a database, a document management system is useless. Content management systems were evaluated: some were too raw, others too complex to be interfaced.

But as this application is fairly specific the architecture is based on a content repository and most end-user functions have been developed to meet precise needs. When an author decides to write a product sheet, the system searches the database in order to identify applicable subjects and generate a default sheet; this default sheet contains items that are propagated from the product family sheet to the proposed product sheet, depending on the content of the database. The author can accept or refuse this proposal, item by item, taking care of course. Other functions, such as sorting contents, re-generating the default sheet and comparing versions of the same product sheets have also been made available to authors.

Processes

Processes are affected by the information model. There are many ways to build them and many available methodologies to do it. But there are also three key issues, which cannot be ignored:

- *Defining the content* The easiest way to define content is to define content types first and the rest (document, information, data and contents) as content type instances, the results of processed content. Defining content types is the modelling phase output. A content type is a structure model and management rules, such as authoring authorization, lifecycle, and so on. For instance, a content type could be a product designation and another one the set of companion products.
- *Designating an owner for each content instance* A content instance can have any owner; an owner is responsible for the life of the instance. A

different type of owner should be designated for each content type, not the same as for the instance, otherwise it rapidly becomes unmanageable. A content type owner cannot be an individual or a department. It has to be a function in the organization, regardless of which department is concerned. But this raises another issue: the organization's structure or who is responsible for which function. In other words the creation of information types may require the implementation of an organization-wide directory, which should be comprehensive enough to handle such information.

- *Defining production and publication cycles* Normally these cycles can be derived from the information model. However, they should be described carefully especially in terms of responsibilities.

So, contrary to what happens in the traditional approach (see the section on modelling information, above), defining processes is clearly a sub-product of the modelling phase. Indeed, it is very close to it.

Conclusion

One cannot build a rule from a single case. However, information modelling is certainly the key success factor in an information management system. Observation shows that many errors arise in organizations systematically, such as:

- *Addressing information by volume throughput instead of its value* Statistics show that a document entering a company is duplicated 12 times on average. Then the 'copy and paste' activity starts and the same document is transformed into hundreds of variants. But it is too late. No one knows any longer which is the original document, so you can throw all of them away. It is much wiser to teach staff how to use basic office automation tools like e-mail messaging properly rather than to search for an ideal technology that could overcome bad practice.
- *Copying existing practice, because it is easy* It takes a lot of time and effort to understand what existing methods are and, until the analysis is complete, there is no change. When the analysis is completed, no one can change anything. That is a real problem. We did not manage information. We managed what computers are able to deal with; remember that they started from scratch. We constantly adapt to their capabilities, which

are still very limited. Human intelligence should drive their progress, not the contrary.

- *Forgetting processes, because they are complicated and also terribly boring* As a designer you have to play the schoolmaster role towards people who do not think they are learning. Rationalization of passionate behaviour is pretty complex. But the principle that works in general physics, which says that any system naturally reaches its balance point, does not work with human beings.
- *Ignoring organizational issues* Changing the rules of writing and asking a marketing person to write structured documents when he or she feels that their day-to-day work is to compose poetic and attractive documents containing boring legal information is a real challenge. The key message that has to be accepted is that everybody works with the objective of furthering organizational excellence, not for their personal pleasure.
- *Selecting software tools, while forgetting the project's objectives* Comparing software tools is a very simple exercise. You just have to ask a thousand questions and rank them. And then you are lost, because it is a useless task. Software tools all provide similar functions and run on similar architecture. And the missing evaluation question is 'What is your content model and how does it fit my content model; does the software allow for it?' This is the only question that matters.

Recommendations:

- Forget the idea of documents: look at content, regardless of media and communication channels.
- One could build the best model in the world, but if no one can understand it and its value, it is useless. Prototypes can be implemented and improved within a short time (a few weeks).
- Do not focus only on existing processes and content. They result from technologies that have been implemented in your company, not from your ideas.

Georges Clemenceau said that 'war is too serious a thing to be left in the hands of military people'. Information is probably too serious a thing to be left in the hands of IT people.