Managing a portal of digital web resources by content syndication

Paul van der Vet (a), Martin Hofmann (b), Theo Huibers (c,a) and Hans Roosendaal (d,a)

(a): Dept. of Computer Science, University of Twente, Enschede
(b): Department of Bioinformatics, Fraunhofer Institute, Algorithms and Scientific Computing Group, Schloss Birlinghoven
(c): KPMG Business Advisory Services, Amstelveen
(d): School of Business, Public Administration and Technology, University of Twente, Enschede

Abstract

As users become more accustomed to continuous Internet access, they will have less patience with the offering of disparate resources. A new generation of portals is being designed that aids users in navigating resource space and in processing the data they retrieved. Such portals offer added value by means of content syndication: the effort to have multiple, federated resources co-operate in order to profit optimally from their synergy. A portal that offers these advantages, however, can only be of lasting value if it is sustainable. We sketch a way to set up and run an organisation that can manage a content syndication portal in a sustainable way.

1. Introduction

The advent of motorways has created a market for one-stop shopping centres. As continuous Internet access becomes more widespread, the distinction in availability between in-house and remote resources loses its significance. The availability of resources thus grows virtually unchecked. However, Internet users can tap the ever-growing plethora of data and knowledge resources available over the web only in principle. Navigation is usually unaided and each resource comes with its own idiosyncratic operating instructions. This situation inhibits the growth of a market for one-stop information services. One-stop information services or portals, as they are often called, aim to provide their users access to information resources in a narrowly defined domain, such as GPCRDB, the portal for information about G-coupled protein receptors. The driving motivation for portals is content syndication: an effort to combine content to provide added value to patrons in making the back office more efficient.

The key success factor for a portal is sustainability. Whatever the portal offers, it should do so with a clear mission, with a clearly defined profile, and with a secured continuity of retrieving it. The current *modus operandi* of many web-based resources and portals is that of self-organisation. It is questionable whether this way sustainability can be assured. In this paper, we want to explore the alternative of an organisation modelled on that of a commercial enterprise for operating a portal in a sustainable way. We present an inventory rather than a complete model and will briefly touch upon a variety of topics to provide a background. The focus is on management and organisation. We will also be dealing exclusively with information produced by the so-called hard sciences like biology and physics.

For the design of one-stop scientific information services, two models stemming from the pre-Web era present themselves: the *repository model* and the *journal model*. They are end points of a continuum rather than models on their own. The repository model is the least ambitious of the two. It views the portal as the WWW analogue of a repository or archive. In this model, the focus is on availability, which in a web environment means navigation in resource space. Like the repository model, the journal model focuses on availability but in addition aims to set a quality standard. Like its source of inspiration, the scientific journal, a journal model portal generally offers less navigation than the repository model and it may cover a narrower field. Because navigation is mandatory when resource space expands, portals that follow the journal model will increasingly add navigation aids, as, indeed, publishers of scientific journals are now providing. The difference between the two models then becomes that of quality assessment. This difference affects the operation of a portal and the possibilities it can offer to its users.

Starting point is that there will be a growing market demand for integration options. Current portals offer access but it is up to the user to further process the information gathered through the portal by means of his own desktop programs, quite a laborious enterprise. Companies have stepped into this market by offering pipelining systems that enable the user to set up a dataflow between applications with minimal effort. Examples of such tools are the Kensington discovery Environment, TurboWorx, and Pipeline Pilot. As such and similar tools become widespread, data taken from resources are increasingly input in complex calculations, so that it is difficult to assess how errors in the data will affect the result of the calculations. Errors in data are unavoidable, however, even when we disregard data entry errors. The data we are considering stem from experimental science that progresses both by new findings and by corrections of old findings that after a while proved to be erroneous. There are large quality differences between resources. Integration thus depends crucially on resources each having at least a predefined minimum quality. In this sense the repository model does not support integration while the journal model does.

In this paper we further explore the issue of portals that follow the journal model by presenting a design for the organisation that sets up and maintains such a portal, in particular for scientific information. Our more specific example will be a fictitious portal for molecular biology. We think that the design can be ported to other scientific domains like materials science, crystallography, or organic chemistry. It seems plausible that the design can also be ported to non-scientific domains, but we have not considered this issue.

The portal has to fulfil a number of technical and organisational desiderata. Among the technical desiderata are:

- A single entry point giving access to a critical number of resources in a homogeneous way.
- Ability to handle resources of different kinds: databases, knowledge bases, and programs, in a predictable way. Because most resources came into being as the result of a relatively isolated effort, operating instructions were and still are re-invented every time and therefore display a bewildering variety. The portal should hide the variety from the view of the user. [1], [2]
- Intuitive navigation in resource space. One way to ensure this is to present the user with an environment familiar to practitioners of the domain, in which resources are accessed by clicking.
- Ability to handle access fees or subscriptions in a transparent manner. Some resources may be free of charge for everyone, others may be fully commercial, and still others may be free of charge for some user groups but commercial for other groups. Pricing schemes, where applicable, may vary. Users should not be bothered with these details but pay the access fee required to a single party.

Organisational desiderata are:

- The portal must be sustainable. This means that there must be an organisation to secure sufficiently stable sources of income at a level allowing its sustainable operation.
- There must be an organised and accounted form of quality control. The quality of resources varies enormously from being indispensable to being a heuristic aid at best. Primacy must be given to active researchers in the field when matters of content are involved, such as quality criteria.
- The portal must also function as a platform for announcing the availability of new resources so that the user is not obliged to rely on information that has to be gained haphazardly.

We believe that by setting up a portal in this way, resources are used more economically and practising scientists can do their work more efficiently. In the following part of this article, we will further explore the organisation that supports this kind of portal by reviewing the four corners of Leavitt’s square, beloved in management science circles: content, process, management, infrastructure. [3]

We will focus on the process and management corners.

2. Content

A portal is of value because it provides access to content that is of interest to a critical number of users. The content fits a profile that can be articulated to such a degree that the portal’s existence and mission can be made known to the relevant communities. The content is typically tied to a particular community. In the scientific disciplines we are considering in the present paper, the content is both produced and used by the same community. Of particular relevance to a portal that adheres to the journal model is the presence of shared quality assessment methods in the user community. By contrast, for a virtual theatre portal, the content producers and consumers constitute different communities. This portal gives access to information about shows, concerts, the main performers, while also being a booking office.[4]

A molecular biology portal will give access to gene databanks, protein and pathway databases, literature abstracts and full-text versions of primary journal articles, sequence alignment tools such as BLAST, and more. As tools become mature, access to programs that perform operations on the data such as pathway simulation software and knowledge bases will be added. The portal presents itself to the biologist as a desktop that enables and supports the complicated operations on data required for research in molecular biology. The portal hides from the user whether resources are in-house, maybe even on the same machine, or remote. Biologists will want to be able to store data they obtain in wet labs through the portal, too, so that seamless integration with other resources is ensured from the start.

An issue related to content is the nature of the quality assessment. The assessment typically relates to entire resources. Items kept by resources will generally have been assessed for quality by the content providers of these resources. As a result, the assessment carried out by the portal should be an assessment of the primary quality assessment process carried out by the content provider. Scientific communities are quite familiar with quality assessments and the conclusions that can be drawn from them. The situation is different, however, in cases where the public is given access to resources. Consider, for example, a hospital that wants to provide access to selected resources for patients and their families. The hospital will obviously not want to warrant the correctness of all items to which it gives access this way. What kind of warrant is implied by the quality assessment procedure of the hospital constitutes a subject for legal and, one may add, moral concern.

3. Process

3.1 Introduction

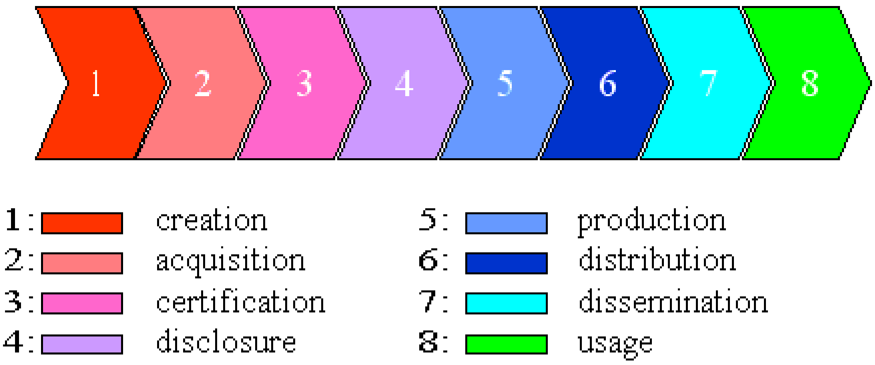
The description of the portal organisation is based on the value chain of scientific information.[5] The value chain consists of steps such that each step adds value to the output of the former step. Each step can be associated with one or more tasks to actually add the value, but the order in which these tasks are performed is only approximately determined by their sequence in the value chain. For example, two values may be added in what is a single process to an institute; or values are added in an iterative process. The entire chain spans a communication process from source to sink.

We use the value chain to define tasks and to allocate them to the various actors that play a role. There are different value chains for different levels of communication; communication may even, at each level, proceed in a different way.

The basic level in the biology example is that of the laboratory, where experiments lead to data that generally are published in peer-reviewed literature. It is possible to discuss the value chain between experiment and refereed paper, but this process is less relevant in the present context and we will regard it as a black box. Increasingly, journals require article authors to deposit their data in a publicly accessible data resource as part of the publishing process. The value chain of a data resource is highly relevant here and we will discuss it below in some detail.

3.2 Value chain of a data resource

The value chain of a data resource can be schematised as in the picture below:



value chain

We will structure the discussion by means of an example that features a fictitious database called E-Base of enzyme properties like chemical structure, 3D shape, genetic origins, and the like. The source of the communication channel is called creation. In the example, it is a black-boxed summary of the laboratory-level processes that lead to publication of enzyme properties in the literature. The acquisition step collects this information from the literature. The certification step subsequently assesses the quality of the data thus gained. In the field of biological databases, this process is often called curation. Note that if E-Base would follow the repository rather than the journal model, this step would consist of a marginal check, for example to ferret out corrupted data. Adding for example metadata in the disclosure step enriches the data for later retrieval. The production step prepares the data for distribution by storing them in a predetermined way on a carrier. The distribution step comprises the digital distribution of the data, including pricing schemes. The dissemination step ensures that the data are disseminated among the appropriate user groups. The end-usage step, finally, constitutes the sink of this value chain.

The value chain is instrumental in organising the tasks that have to be done in order to bring the contents of E-base to its users because, with the obvious exclusion of the creation and end-usage values, the addition of all other values corresponds to identifiable tasks. The end-usage value constitutes the *raison d’être* of the organisation that maintains E-base. One of the discussion points is who should do what tasks. Currently, it is not uncommon to see that an organisation like the one that maintains E-Base performs every task in-house.

3.3 Value chain of a portal organisation

The key observation that underlies the design of our portal is that precisely the same value chain can be assumed at the meta-level of an entry point that provides access to resources. The units transmitted this time are not data but entire resources. Thus, the creation step refers to the process of creating and maintaining resources available over the Web. From the point of view of the portal, this is a black box. In an acquisition step, the organisation responsible for the portal selects candidate resources for addition to the resources it makes accessible. This involves acquisition of a URL and negotiations about conditions of use such as price. The portal wants to be able to give its users an indication of the quality of the resource or, more generally, their 'value for money', whatever the currency may turn out to be. Quality judgements are produced in a certification step. Obviously, in this case the tasks that correspond to the acquisition and certification steps are closely connected because the quality of a resource is an important factor when the portal organisation determines whether it wants to add access to the resource and, if so, at which price. A review committee consisting of domain experts will provide guidelines on the acquisition of resources and their quality level.

In the disclosure step, the portal organisation adds meta-data such as annotations, cross-references, and navigation aids to the resources in order to prepare for easy access by the end-users. The actual work of adding the annotations is done in the production step. The value of the production step is added in two ways: providing the actual access to the resource (by a hyperlink, by mirroring, or in another way) and by ensuring interoperability of the data stemming from different resources.

Addition of the distribution value again involves two tasks. Physical distribution is implemented by means of known server technology. The other task associated with distribution is that of pricing and marketing. Adding the values to the resources by the portal organisation will inevitably incur costs. Adequate funding has to be found for the portal organisation, either as public funding or direct funding by charging the customers, or a combination thereof. A possible scheme could offer two versions: a minimal version at a low charge or free of charge, provided the funding allows this, and a 'de luxe’ version that comes at an additional price. The pricing scheme may involve more modalities, however. The use of some resources will no doubt involve fees. To make matters even more complicated; some users of the portal may already have a subscription to some other resources and do not want to be billed twice. This means that issues of pricing and marketing are an important concern.

The addition of aids for end-user navigation is the main value added by the dissemination step. We believe one attractive option is to allow the user to travel in an environment that portrays the scientific domain. Unlike what is the case in traditional virtual reality, the idea is not to mimick reality as closely as possible. Rather, the visualisations help the user to navigate in resource space by making the required distinctions and showing the important relations in a visual way. Finally, a part of the dissemination value can also be added by a client, such as an institute that wants its own, proprietary data accessed together with other resources through the same interface.

End-usage, finally, is within the scope of the portal organisation insofar as expectations of the kind of end-users and their working practices and needs of course drive the entire design.

4. Management and organisation

Some organisation must run the portal and assume overall responsibility for its proper operation. This organisation should be held accountable for the processes outlined above. This organisation should be able to guarantee its stakeholders sustainable utilisation of the portal and the knowledge available and accessible through the system. A portal federating a number of resources allows a lean organisation. This organisation will be faced with a number of strategic and operational objectives.

There are two main strategic tasks to be performed. A most crucial task is to represent the full international community of users and creators of knowledge sources in the project. This is the representation task. This task can best be fulfilled at two levels. At the highest organisational level there is a senior international representation of the entire community. At the operational level, we envisage user groups that meet regularly. Furthermore, the organisation should be able to develop and implement a clear strategy based on the above meta-level value chain for the portal. This is the executive task. The executive task comprises overall responsibility in managing the portal and laying down and deciding on the overall strategic framework for the tasks.

The portal organisation should be able able to achieve the following strategic and operational objectives:

1. Representing the full international community of users and creators of knowledge sources in the fields of interest for the project. This task should be fulfilled at two levels. At the highest organisational level there is a Supervisory Board (SB) that consists of a senior international representation of the entire community. At the operational level, we envisage user groups that meet regularly.
2. Being able to further develop and implement a clear strategy based on the above meta-level value chain for a federated knowledge ensemble for the project. This is the executive task that is entrusted to a small Executive Board (EB). The EB is hired and fired by the SB. The EB manages the consortium and lays down and decides on the overall strategic framework for the tasks mentioned here. To give an example: the EB sets after due consultations the general conditions for certification of resources, while the certification itsef, including a decision on the admission of a resource to the ensemble, is delegated to another body (see objective nr. 7 below).
3. Being able to operate as an international organisation. A task for the EB.
4. Being able to protect the interests of the ensemble such as but not limited to property rights and sustainable continuity of the services in the future. A task for the EB.
5. Being able to provide conditions furthering the sustainability of the participating resources. A task for the EB.
6. Being able to contact and negotiate with suppliers of these resources. This task is best performed by a small acquisition team (that, of course, reports to the EB and acts upon guidelines issued by the EB).
7. Being able to assess the quality of these resources in an independent way. This is the task of an international Review Board (RB). The RB is the ultimate authority in the organisation to approve of the admission of resources into the ensemble and is composed of international experts in the field. The RB is appointed by the EB and performs its task on the basis of a set of formal certification rules as laid down by the EB. It needs to be seen if the RB should possibly be divided into divisional RB's (DRB) to represent a finer granularity of the different subject fields.
8. Being able to warrant the intellectual integrity of these resources. A task for the EB.
9. Being able to market and sell the ensemble under conditions to be agreed. This is the task of a marketing and sales team reporting to the EB and performing its task on the basis of a set of marketing and sales rules as laid down by the EB.
10. Being able to ensure optimal interoperability between the participating resources, with consequences for the interaction between creators and participating resources, and between users and resources. This is the task of an international Standards Committee (SC). The SC is the ultimate authority in the consortium to ensure optimal interoperability of the resources present in the ensemble and is composed of international experts in the field. The SC is appointed by the EB and performs its task on the basis of a set of formal standardisation rules as laid down by the EB. The SC sets guidelines in the following areas of production: data exchange/XML, data management issues, ontologies, orthology, and other areas.
11. Being able to materially create the ensemble, and to operate, maintain, update and expand the ensemble. This is the task of the production team (PT), possibly comprising of a number of specialised

divisions, reporting to the EB and performing its task on the basis of a set of production rules as laid down by the EB. The production team or one of its divisions is supported by an set of expert groups to realise implementation that guarantees interoperability and error-free distribution.
12. Being able to give support to the users (international helpdesk). In particular for those users who will make of the ensemble for purposes representing high risks, such as in clinics, complicated research set-ups, in connection to patients, etc. This task requires a helpdesk.

An organisation as sketched above will be able to operate the portal in a sustainable way that may count on adhesion from the majority of users. We are convinced that there is a market to warrant the investments needed to realise the portal.

5. Infrastructure

Realisation of the portal is largely possible with existing technology. The main technical decision is whether to design the system of portal and resources as a data warehouse or as a federated information system. The pros and cons of either solution are well-known and can be briefly summarised here. A data warehouse gives guaranteed access to all resources and can guarantee interoperability. Also, a data warehouse can be shielded from the outside world except during the brief intervals in which new data and/or resources are added. Against this, maintenance of a data warehouse constitutes a huge and, for many scientific user communities, prohibitive effort. For institutes that can afford the expenditure, a data warehouse is probably the best solution. Indeed, large pharmaceutical and agrotechnical companies routinely establish data warehouses for their in-house researchers, if only because this way, confidentiality of the data and findings can be safeguarded.

A federated information system, [6] by contrast, is an open environment. Maintenance of resources is left to the groups that make the resource available. Maintenance costs for the portal comprise the implementation and maintenance of middleware, of the navigation interface, and of the interoperability layer. Against this, a federated information system relies on a complex configuration of often implicit agreements. For example, resource providers are required to operate their resource in a predictable way, meaning, among other things, to have their data available round the clock and to deliver their data in a format of which the syntax may be unique to the resource but is always known and the semantics is agreed.[7] It is one of the tasks of the portal organisation to make the necessary agreements explicit. Navigation and interoperability are aided by making use of existing consistent semantics and adding semantics where needed. For biology, this semantic interoperability is served by the Open Biology Ontologies initiative. Portals are considered by such diverse organisations as E-BioSci, ORIEL, and BioASP.

6. Further outlooks

The portal organisation will quite naturally assume other activities in fulfilling its mission as general clearing house for information in the chosen domain or domains.

A natural extension of its tasks is to commission literature reviews and other compilations of a predefined quality level. These compilations are in turn available as resources, i.e. via the graphical interface. More importantly, they are structured using meta-data standards and other guidelines, and they can be heavily linked to other resources. This kind of reviews then far surpasses more traditional kinds in terms of reader value.

The developed standardisation products can be tools for a more disciplined data management and experiment description or annotation than is customary today. An important task for the portal organisation in the biology domain will be to bring together existing ontologies and ontologies that will have to be developed so as to span the entire range from molecules to populations, over molecular complexes, organelles, cells, tissues, organs, body parts, and organisms.

Somewhat further in the future lays the use of consistent semantics developed by the organisation, such as in biology ontologies. Consistent semantics structure content and therefore are important didactical aids. They can also be used as a scaffold for constructing a knowledge representation of a major part of a scientific paper. Specialised authorware would construct the knowledge representation in a way that is transparent to the author. For readers, a knowledge representation enables personalisation of the article.

7. Concluding remarks

Resources multiply every day. They are hard to find and their operation requires knowledge of ideosyncratic instructions for use. User communities depending on the availability of resources waste time and money in collecting and processing data, quite aside from the real possibility of errors creeping into and propagating throughout the system. The disadvantages of this state of affairs are now becoming apparent to a number of user communities. These communities are actively seeking ways to remedy the situation. Often, however, the remedy takes the form of a "roll your own"-portal that is operated with uncertain future by one group, while another group with different ideas offers a portal with an equally uncertain lifetime but divergent operation. This way, the advantages of content syndication are not fully exploited and the diversity of resources is simply echoed at a higher level of aggregation. In science, user communities can start scholarly journals, so there is no reason why they could not also start an organisation whose purpose it is to establish and operate a portal in a sustainable way. For the examples we have considered the organisation is international and will almost inevitably be world-wide.

Portal organisations have a vital role to play in scientific research. They can fulfill this role if managed properly, by an organisation that ensures sustainability and assigns responsibilities where they belong.

References

1. P.E. van der Vet, "Building web resources for natural scientists", in: *Interactive distributed multimedia systems and telecommunication services (IDMS2000)*, H. Scholten, M.J. van Sinderen (eds.), LNCS 1905, Berlin: Springer, 2000, pp. 205-210.
2. L. Stein, "Creating a bioinformatics nation", *Nature* 417 (2003) 119-120.
3. A.V. Malchanau, P.E. van der Vet, H.E. Roosendaal, "Habitable Interfaces: an approach to federating information resources for scientific communication", *submitted*.
4. H.J. Leavitt, "Applied organisational change in industry: Structural, technological, and humanistic approaches", in: *Handbook of Organisations*, J. March (ed), Chicago: Rand McNally and Co, 1965, pp. 1144-1170.
5. A. Nijholt, J. Hulstijn, A. van Hessen, "Speech and language interactions in a web theatre environment", in: *Proceedings of the ESCA workshop on Interaction Dialogue in Multi-Modal Systems*, P. Dalsgaard, C.-H. Lee, P. Heisterkamp, R. Cole (eds.), Aalborg: ESCA/Center for PersonKummunikation, 1999, pp. 129-132.
6. H.E. Roosendaal, P.A.T.M. Geurts, "Scientific communication and its relevance to research policy", *Scientometrics* 44 (1999) 507-519.
7. P.M.D. Gray, G.J.L. Kemp, "Federated database technology for data integration: lessons from bioinformatics", in: *Electronic collaboration in science*, S.H. Koslow, M.F. Huerta (eds.), Mahwah NJ: Lawrence Erlbaum, 2000, pp. 45-72.
8. P.E. van der Vet, H.E. Roosendaal, P.A.T.M. Geurts, "C2M: configurable chemical middleware", *Comparative and Functional Genomics* 2 (2001) 371-375.