

A Regional Automated Data System for the Distributed Publication Environment.

A concept paper of the SouthWest Data Center, Inc.

This Request for Comments¹ explores the requirements for developing an integrated and maintained regional GIS resource from multiple and disparate production sources. The Goal Statement at the Colorado Plateau Data Coordination Group Workshop (Farmington New Mexico on October 28 – 29, 1997) serves as the impetus for this particular exercise.

“Our goal is to improve our ability to effectively collect, manage, and transfer data across all political boundaries enabling more improved decisions to be made regarding the Colorado Plateau region.”

This language conveys a need to aggregate a large number of data sets pertaining to a region that spans five states. Each state contains numerous jurisdictional entities and, their boundaries define various federal agency regions. The inclusion of multiple sovereign Indian nations, some of which abut or cross state lines, add to the complexity. The intention of the Colorado Plateau Data Coordination Group (CPDCG) is to “publish” these aggregations as the Resource Atlas of the Colorado Plateau.

Since its creation, the CPDCG has had a difficult time “putting its arms around” this task. For this reason the SouthWest Data Center, Inc. conducted an inventory of Internet accessible administrative, political and land ownership boundary data sets in the five state region. The purpose of this boundary inventory was to gain an understanding of the larger challenges by focusing on one component – the inventory of data sets for one core data (framework) layer - boundaries.

Colorado Plateau Boundary Integration Project

Using various means for locating GIS data on the Internet, 813 data sets were located, downloaded, re-projected and organized into a directory based on scale. These data sets are available via FTP at <ftp://ftp.landuse.com/GeoAtlas/>. An Access 2000 database which cross links the re-projected data sets to their source can be found in the /GeoAtlas/boundary folder. The [Colorado Plateau Boundary Integration Project](#)² report describes the methodology, illustrates the significant problems encountered and makes recommendations for further course of action. The report was presented at the October of 2001 South West Users Group (SWUG) convention in Tucson, Arizona.

Prior to the boundary inventory several problems were known to exist in combining data from disparate sources. Among these problems were the lack of attribute standards and the use of dissimilar boundary lines. Less apparent problems encountered were the complete dissimilarity of naming conventions, or classifications used, the diverse protocols required to access the data via the Internet and intermittent availability of projection information. While the lack of standards and common boundaries impair the aggregation process, the obstacles found in the gathering process make integration and maintenance of disparately produced data sets financially

¹ Comments are welcome at rcm@landuse.com

² http://www.landuse.com/colorado_plateau/boundary_project/

impractical. Until other data gathering means can be developed the Resource Atlas of the Colorado Plateau will remain a topic for discussion and not a serious endeavor.

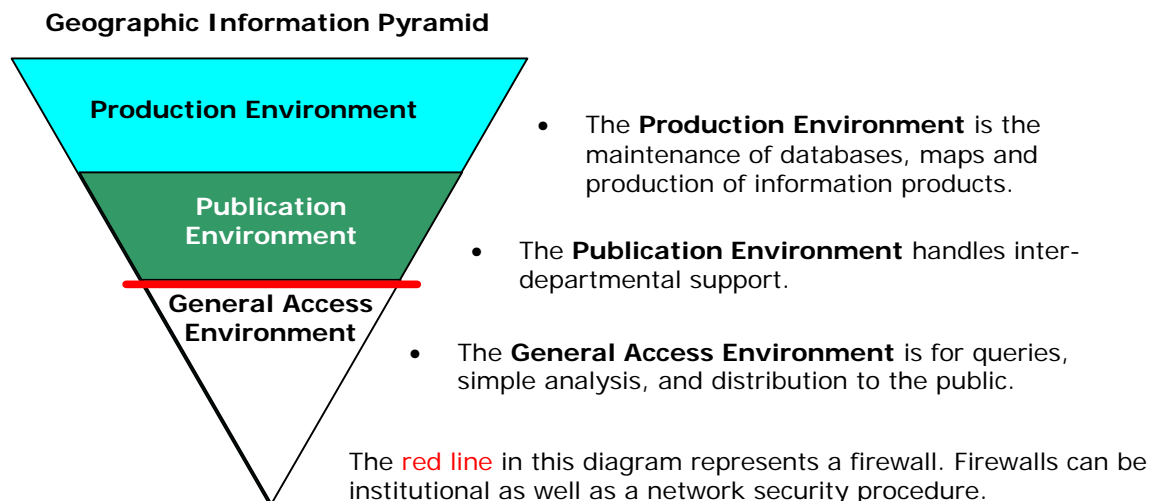
Regional Automated Data System

The fundamental weakness of the current gathering process is its reliance on a “pull-process”. In this case the “pull-process” uses manual labor to find and process the datasets into a usable form. To overcome the manual labor component of gathering and processing data sets, a “push-process” must be used instead. The “push-process” allows disparate data producers to contribute information and data sets to an automated system for integration. This automated system allows data producers to continue using their own conventions while also contributing key information into an integrated format. The integrated format can then be used by end-users to search, view and retrieve data sets efficiently. By duplicating this automated system on other Internet servers through out the Colorado Plateau region the CPDCG can then, as their goal states, “effectively collect, manage, and transfer data across all political boundaries” using a distributed network. This “push-process” can be achieved using a **Regional Automated Data System (RADsys)** in a **Distributed Publication Environment**.

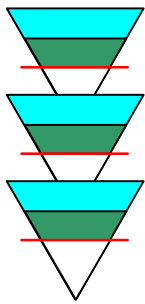
The Geographic Information Pyramid

The relationship between a **RADsys** and a **Distributed Publication Environment** is best understood by first looking at the Geographic Information Pyramid as presented in [Production, Analysis and Publication A Concept for Geographic Information Environments](#)³. Pertaining primarily to Cadastral data, the underlying concept is also applicable to other Core Data Sets (a.k.a. Framework Layers) including boundaries. For the purposes of this document the diagram has been inverted.

The proposition: within every disparate Political and/or Geographical constituency there are three basic conditions or states for geographic information, with each condition being represented as an environment.



³ Production and Publication A Concept for Geographic Information Environments - Nancy von Meyer, Scott Oppmann, Norm Bushor, Chris Lucas (<http://www.fairview-industries.com/articles.html>)

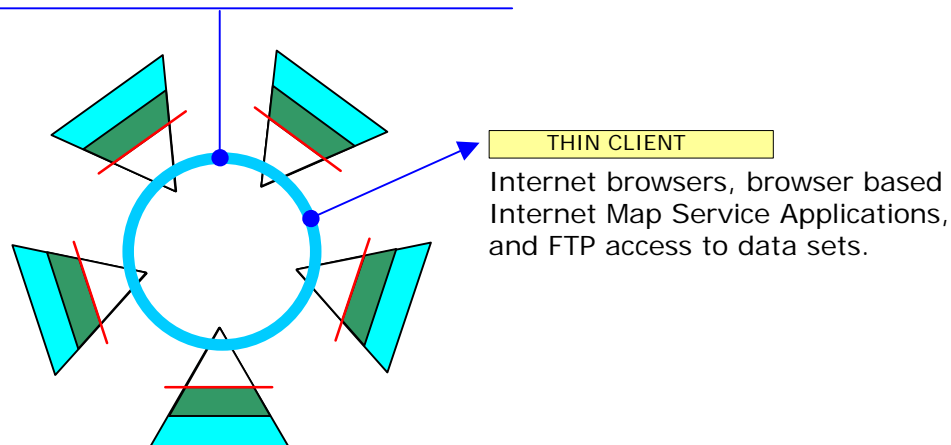


Within each Political and/or Geographical constituency, **information flows** in one direction – from the Production Environment to the Publication Environment to the General Access Environment. Between Political and/or Geographical constituencies however, information can, and does, flow from one entity’s General Access Environment into another’s Production Environment. Once data sets are made available for general access they are often incorporated into the production of new data sets by an entirely disparate data producer. The transfer of information between **Jurisdictional Publication Environments** is made possible using two means; 1.) physical storage devices, and 2.) the Internet.

Transferring information via physical storage devices, such as CDs, requires the manual process of copying the information and the need to physically deliver the storage device using some means of ground transportation.

Transferring information via the Internet can be achieved using view-only Internet map services or FTP delivery of actual data sets.

Networked Jurisdictional Publication Environment



The use of the Internet, in most cases, is the preferred means for transferring information between constituencies as it can be transferred more efficiently with reduced burdens on constituency staff. However, as found in the boundary inventory, this Networked Jurisdictional Publication Environment does not enable a practical means for a regional integration of data. As stated earlier, FTP access requires any number of protocols that are unique to each Political and/or Geographical constituency. Furthermore, naming conventions also remain unique to each constituency. The process of collecting data in the Networked Jurisdictional Publication Environment is extremely labor intensive and keeping data current is nearly impossible.

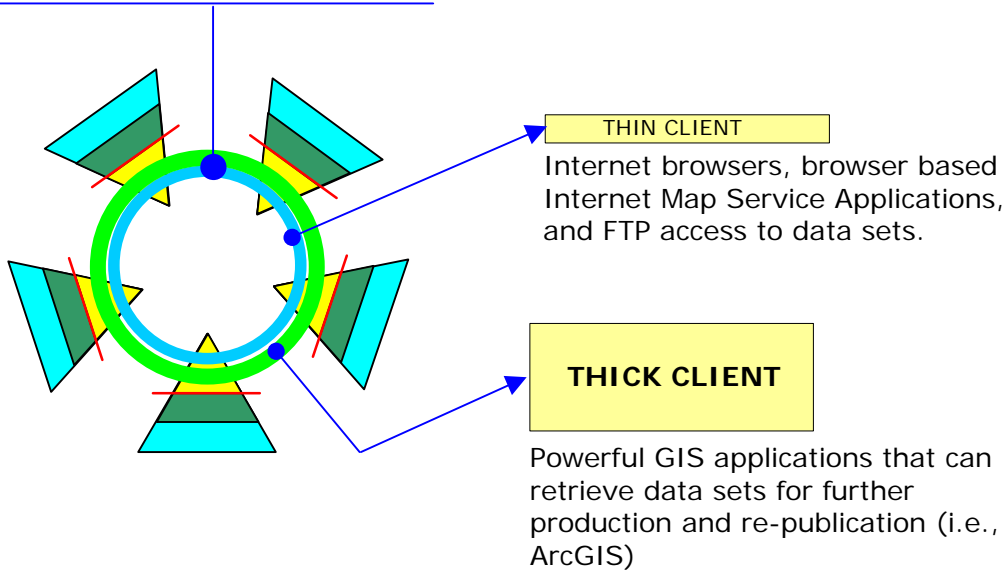
The Geodatabase Server and the Distributed Publication Environment

Using a **Geodatabase Server** in a **Distributed Publication Environment** will partially solve the data-gathering problem.

A Geodatabase Server is:

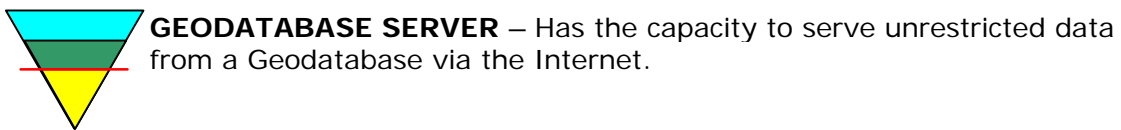
1. An Internet data server
2. Versioning capable
3. Rules capable
4. Open GIS standards compliant.

Distributed Publication Environment

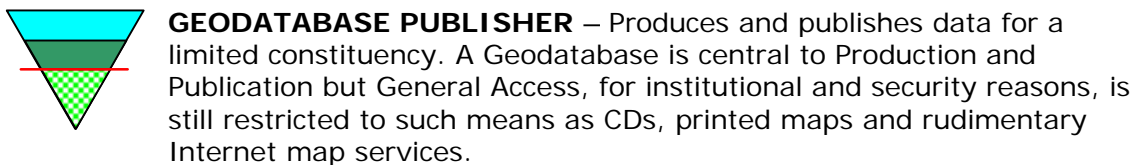


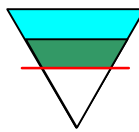
Geodatabase Servers enable various **Thick Client** applications to access unrestricted data via the Internet. This solves the problem with unique protocols being used for data access via the Internet. The problem with disparate naming conventions can also be partially solved by implementing crosswalk strategies between Geodatabase Servers. The obvious hitch in this solution is the disparity between the “**Haves**” – access to a Geodatabase Server – and the “**Have-nots**” – no access.

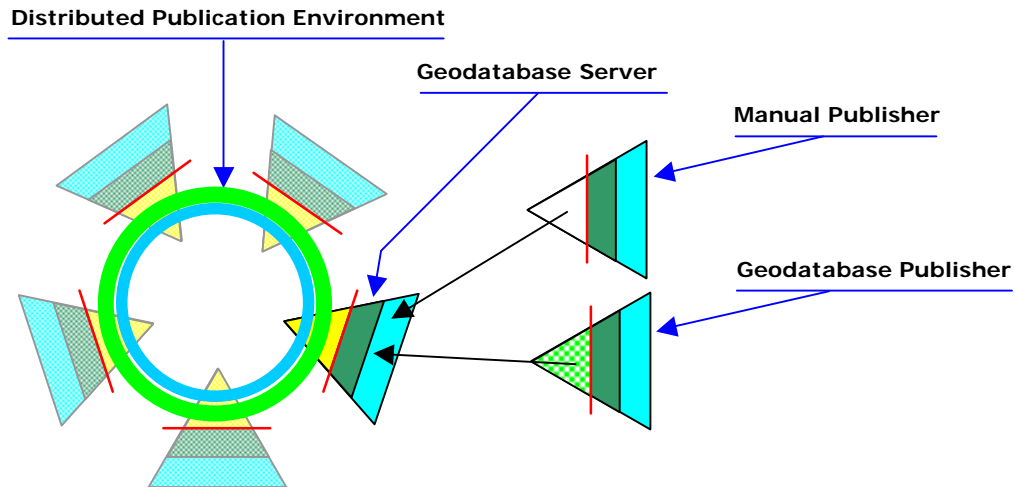
The Haves are those Political and/or Geographical constituencies that use Geodatabase Servers.



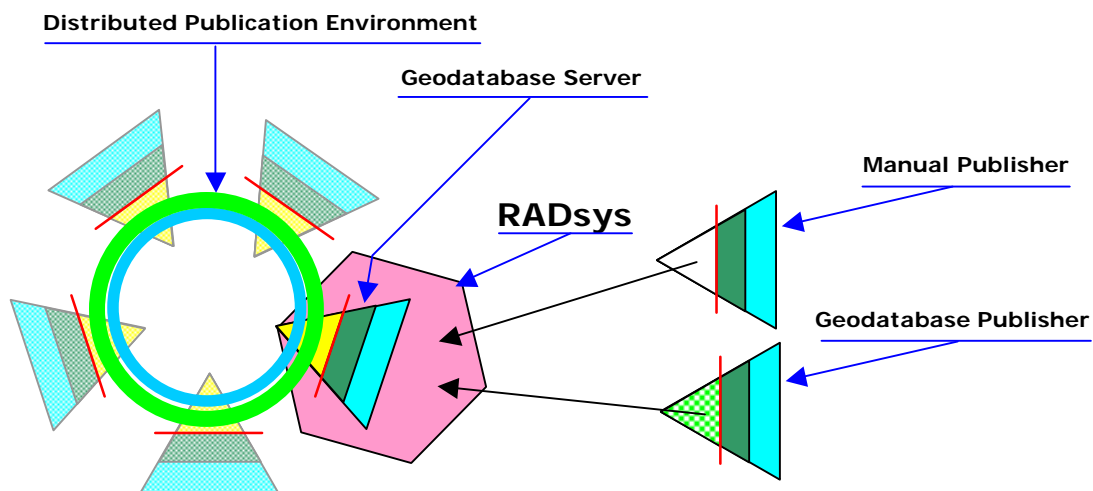
The Have-nots are a larger group of Political and/or Geographical constituencies who do not have access to Geodatabase Servers, and are referred to as geodatabase and manual publishers.



 **MANUAL PUBLISHER** - Produces and publishes data using desktop applications. General Access is made through means other than a Geodatabase server (i.e., FTP sites, CDs, printed maps, Internet Map Services).



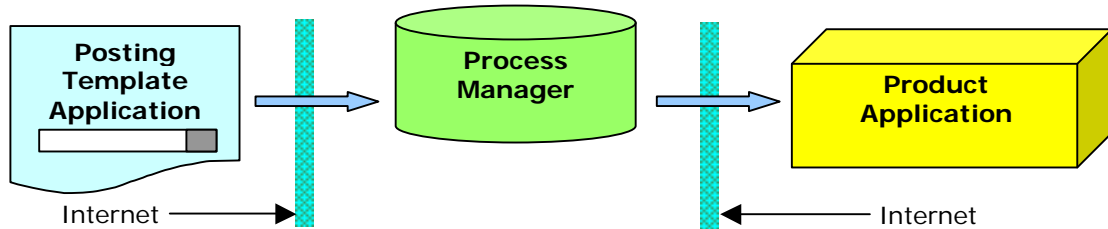
An **information flow** model in the **Distributed Publication Environment** works much the same way it does in the **Jurisdictional Publication Environment** - one constituency's General Access Environment flows into another's Production Environment. This allows Have-nots data a path to feed information into a Haves Geodatabase Server. However, this does not address the problems with the manual integration process. The continued reliance for manual processing does not address the continued need to keep data current. The Regional Automated Data System (RADsys) is needed to solve this problem.



RADsys has three components:

1. **Posting Template Application** - An Internet browser based application for posting data sets, metadata and user input describing this information.
2. **Process Manager** - A Spatial Data Engine (SDE)/ArcObjects enhanced RDBMS (Relational Data Base Management System) for managing the collected data.

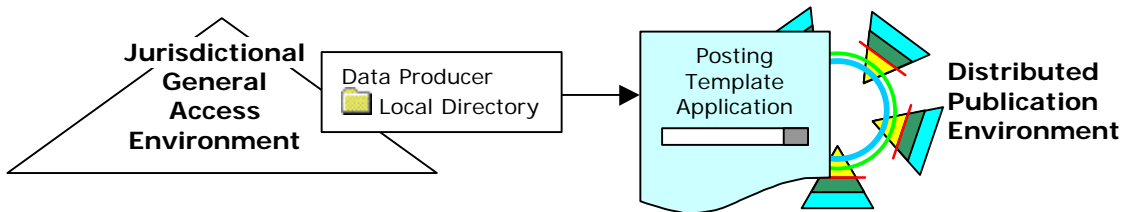
3. **Product Applications** - A set of Internet browser based applications connected to the Process Manager, via the Internet, that facilitate the serving of information and data.



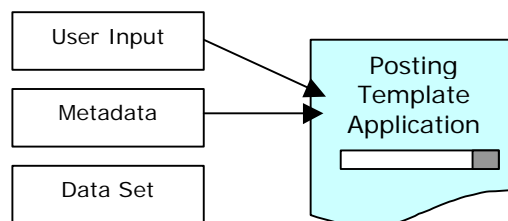
The RADsys overcomes the fundamental problem of data integration and maintenance in cross-jurisdictional environments by allowing disparate data producers a means to easily contribute their data (Posting Template Application) to a data pool (Process Manager) that is part of a Geodatabase Server.

The Input

It is necessary to allow data producers to use their own unique **Jurisdictional General Access Environment** preferences (such as classification names) and also conform to a **Distributed Publication Environment's** conventions. The **Posting Template Application** allows these two conditions to co-exist.

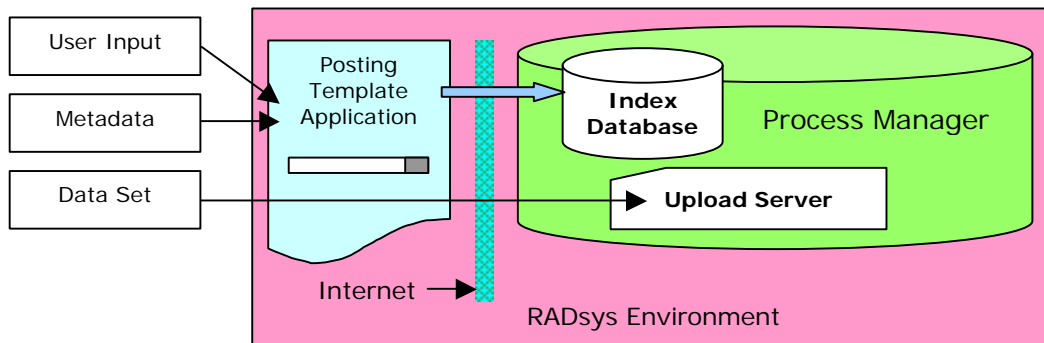


The **Posting Template Application** is a **Thin Client** (browser based) means for putting crucial information, including metadata and its projection sub-set, into the server side **Process Manager**. Although data producers with FGDC compliant metadata can import much of the appropriate information into the Posting Template Application, indexing information, such as classification names, will require using a fixed set of pre-selected information. For example, in the data pool of the Process Manager, all data sets pertaining to administrative boundaries will be classified with a common name. (For the purposes of the Resource Atlas this fixed information will reflect the "Framework" layers plus the others identified as critical – such as geology, soils, etc.)



The user will also indicate, from a fixed set of selections, which data compression application they are using.

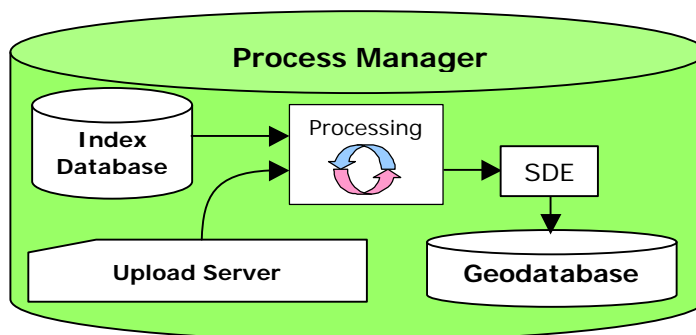
Once the Posting Template Application is completed, the data and attached information can be “published” to the Process Manager via the Internet.



When a data set is “published”, information is transferred to two places in the Process Manager:

1. The User Input and Metadata information, obtained through the Posting Template Application, is transferred to a relational database designed to manage the information used to integrate and distribute the data sets - the **Index Database**.
2. The data sets, as “transfer files”, are sent to a temporary location within the Process Manager – the **Upload Server**.

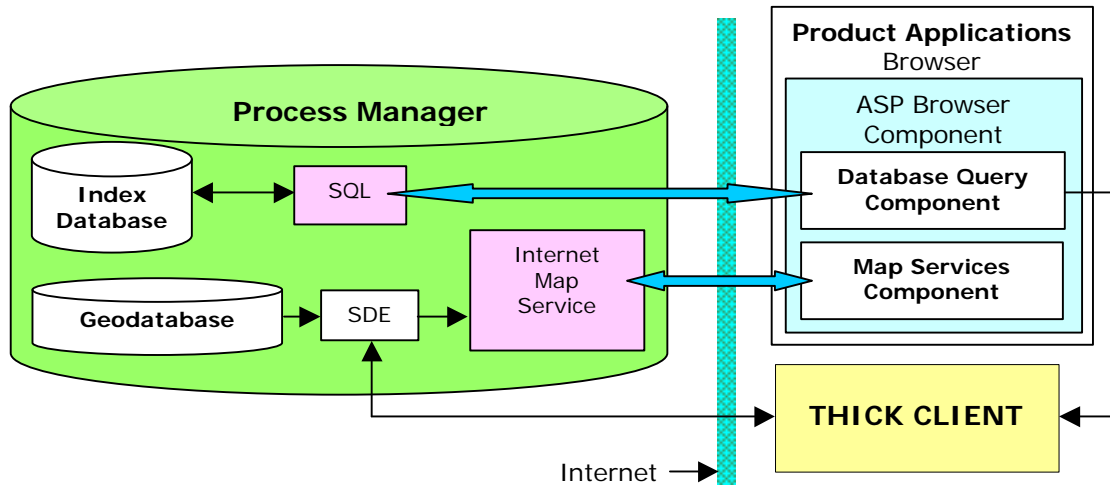
The Processing



Data sets transferred to the **Upload Server** are first uncompressed. Utilizing classification and projection information from the Index Database the data sets are re-projected – using ArcObjects and SDE – into the **Geodatabase**. It should be noted that although an “original” data set (and metadata) exists, and can be hyperlinked to the “daughter” data set processed to the Geodatabase is new and thus has its own metadata. The Z39.50 protocol can be used to search this metadata.

The Output

Up to this point the discussion has been limited to getting the data into the system and maintaining it. The discussion will now focus on how the data can be made useful.



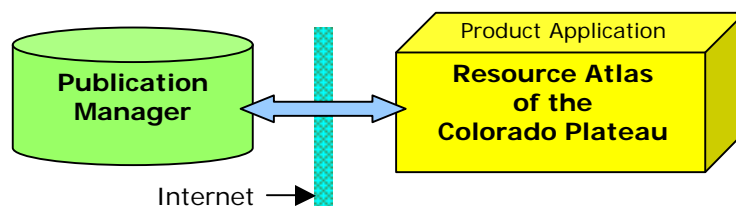
The two primary components to “presenting data” are:

1. **Structured Query Language** – Through SQL the Index Database becomes a hyper “Card Catalog” and “Search Engine” for referencing and prioritizing information about the data sets and their relationships.
2. **Internet Map Service (IMS)**. The IMS is a powerful means to reference data sets and their relationship in a spatial context.

Active Server Pages (ASP) allow SQL and IMS to interact on the server side (where power and speed can be optimized) and efficiently deliver a product to an end-user via an Internet browser.

Product Applications refers to a family of Internet browser based applications that use information derived from the Process Manager. The utility of each specific application is dependent on the “customers” for which it is developed. Where the Colorado Plateau Data Sharing Group may have one set of needs (Resource Atlas), local governments have another (Map Books). All of the “customers” work off of the same Process Manager using pre-defined operations embedded in their particular application. Common to all Product Applications is their ability to leverage a SQL query into a spatial context using Active Server Pages with Internet map services. Spatial queries can work vice versa to extract information from the Index Database using the IMS. Advanced users can always use straight SQL to obtain their own “custom” results. Users with Thick Client applications can connect directly to the **Geodatabase Server** via the Internet.

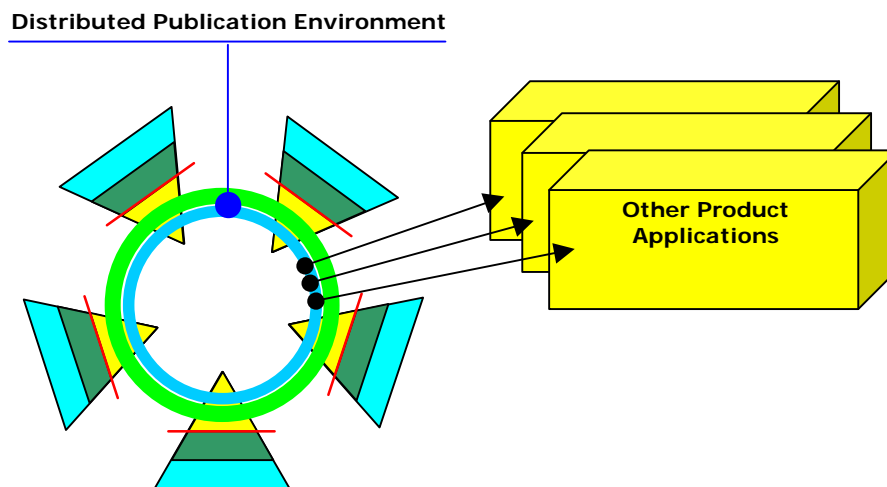
The Products



- **The Resource Atlas of the Colorado Plateau:** Conceived by the Colorado Plateau Data Coordination Group, the idea is still amorphous. However, drawing from the goal of “improving an ability to effectively collect, manage, and transfer data across all political boundaries” certain product needs can be identified.

These product needs are:

- ▶ The ability to bring together like type data sets into a common view for the purpose of illustrating the disparities a spatial context.
 - Utilizing the common classification names now located in the Data Index pre-defined queries can create “virtually aggregated” views of the selected layers. These views will illustrate, in a spatial context, the disparities in boundary lines and attribute types.
- ▶ A context for identifying stakeholders and facilitating the discussion about standards and share codes for same-type data sets.
 - Problems discovered in the spatial context can be linked back to the data producer via the Index Database.
 - Adjoining data producers can communicate directly to resolve minor issues of common boundary lines.
 - “Accepted” standards can be made available to data producers for use in future data production.
 - RFPs can include clauses requiring the use of these “accepted” standards in contract work.
- ▶ An I-Team⁴ tool for recognizing the condition of available data layers.
 - Map views and analogous reports can be generated regarding the availability, date of production and production sources of same-type data sets.
 - Understanding where gaps occur will enable the group to focus in a coordinated effort to rectify these problems.



⁴ The I-Team concept is a product of the Office of Management and Budget (OMB). The purpose of I-Teams is to identify the condition of various layers of data and develop a plan for prioritizing the completion of sets. Through I-Team Reports, the OMB can help direct federal funding to these efforts. Typically I-Teams are implemented on a state basis. However, there is much interest in developing regional I-Teams also.

The Resource Atlas for the Colorado Plateau is only one of many products that can be developed using information available from the RADsys enhanced Distributed Publication Environment. Other products might include:

- **Bureau of land Management Inventory** – Utilizing the “off-site” status of the RADsys/Geodatabase Server, BLM data producers would be able to post and maintain their data sets to an “un-official” clearinghouse for the purposes of understanding various integration problems in a spatial context - without the “liability” of officially endorsing the information. Hyperlinks to the “official” data sets would be option left to the BLMs discretion.

- **USGS National Map – Finer than 1:100,000 scale** – The RADsys/Geodatabase Server, when deployed in other regional areas, will enable the posting and maintenance of data sets not included in the scope of the National Map. Connections can be made to the Geodatabase Servers and the National Map Data Server concurrently.

Creating usable products is the key to the RADsys’ success and subsequent utility. As the products gain acceptance among users, data producers will not only become more inclined to want their data available but to also keep it current. With shared data as the “whole product,” data sets contributed for one product will be available for other products. Of course security protocols can be implemented to make particular data sets available to only appropriate users.

Conclusion

The Colorado Plateau Data Coordination Group has tried valiantly for the past four and a half years to “put its arms around” the task of collecting, managing, and transferring data across the political boundaries of the Colorado Plateau region. Aside from a set of course level data (1:2M) on CD’s⁵ and the boundary inventory produced by SWDC there has been little progress.

As a result of the boundary inventory conducted by SouthWest Data Center, Inc., several fundamental problems in integrating data from disparate sources have been identified. These problems stem from each Political and/or Geographic constituency’s use of unique naming conventions and access protocols coupled with intermittent projection information. In using the conventional “pull-process” for gathering data, and keeping it current, is unrealistic for the purposes of integrating data across a large regional area such as the Colorado Plateau. For this reason a “push-process” needs to be developed. The **RADsys /Geodatabase Server** in a **Distributed Publishing Environment** embodies such a “push-process”.

The **RADsys /Geodatabase Server** in a **Distributed Publishing Environment** model allows for Political and/or Geographic constituency’s that do not use Geodatabase Servers a way to easily contribute and maintain information for regional projects such as the **Resource Atlas for the Colorado Plateau**.

Furthermore, the **Distributed Publishing Environment**, as a de-centralized model, solves the unmentioned problems of “centralized” depositories and clearinghouses. A de-centralized model is the only viable to means for accommodating the potentially vast amount of data storage required for the **Resource Atlas for the Colorado Plateau** to succeed.

⁵ The data from these CDs is available through Utah’s AGRC’s FTP site <http://agrc.utah.gov/colorado.htm>