

# Federal Spending Transparency: Unlocking the Power of Abstraction

Presented to OMBWatch's  
Conference to Develop a Long Range Vision for Federal Spending Transparency  
May 20, 2010

Jim Harper  
Director of Information Policy Studies, The Cato Institute  
and Webmaster, WashingtonWatch.com

Abstraction is the idea of considering the general characteristics of things apart from concrete, specific instances. Abstraction is a powerful tool for comprehending our world—and for remaking it.

Consider the abstractions that we use every day, things like language, numbers, and time. Language is the use of uniform sounds to describe things. The invention of language (and the logic that underlies it) is a big part of what separates humans from other members of the animal kingdom. Writing is a further abstraction on language that improved humans' abilities to catalog, comprehend, and improve their lives.

Numbering is a similar abstraction that allows people to measure and compare quantities of things, a huge innovation from early in human history. Without it, life would be unbearable by today's standards.

Time is another abstraction. It describes the movement of the earth with respect to the sun. The uniform system we use for describing "time" lets people synchronize their activities with others, a tremendous aid to living in an organized society.

More recent advances in abstraction show how it continues to improve our ability to work with the world around us. Expressing letters, numbers, sights, sounds, and symbols as 1s and 0s—digitization—is a way of abstracting information that was central to the invention of computing and the Internet. Without digitization, us policy wonks would still be waiting for Xerox copies of bills to be delivered to us by mail or courier.

The Internet protocol (TCP/IP) is an abstraction that rode on the invention of digitization. It's an abstract way for computers to talk to one another. The Internet protocol paved the way for html (hypertext markup language) and the World Wide Web. These invented abstractions deliver information in readily usable form to computers and connected devices across the globe.

Abstraction is powerful because it allows people to work together, using agreed standards for communicating information, to solve the problems in their lives. It can definitely help solve problems in the area of federal spending.

Earlier this year, a small group of earmark transparency activists put together an abstract model for describing earmarks. Our work is presented at [Earmarkdata.org/schema](http://Earmarkdata.org/schema) and it is attached to this paper as an appendix. The goal, simply put, is to get information about earmarks from Congress in an abstract, and thus useful, form.

Getting earmark data will allow web sites, researchers, reporters, political scientists, and the public to manipulate and use earmark information in any way they choose. People will be able to learn more about whatever they are interested in: They will be able to more easily compare earmark requests and awards with campaign contributions, political party, or seniority, for example. They will be able to make whatever arguments they want to about particular earmarks, earmarking processes, or the practice of earmarking itself. The great thing about abstraction is that it permits new and innovative uses of data—uses we won't know about in advance.

The earmark data model consists of three basic elements:

- **Entities:** “Entities” are *things*. An entity might be an elected official, a bill, or specific provision of a bill. Agencies, programs, contractors, and grantees are all likely entities to describe in a data model.
- **Properties:** Entities are made up of a collection of *properties*—the characteristics that make an entity what it is. The properties of an elected official entity would include things like first name, last name, state and district represented, and so on. One important property of a bill is its bill number (e.g., H.R. 123, S. 1020). Other properties of bills might be the committee(s) they are referred to, their stage in the legislative process, their texts at different stages, and vote tallies on them.
- **Entity-Properties:** An entity-property is a characteristic—a property—that is another defined entity. When a bill (an entity) has as a property that a particular elected official (also an entity) co-sponsored it, that's an entity-property of the bill. (Likewise, the bills an elected official co-sponsors can be entity-properties of the official.)

Entity-properties allow data users to weave together the “stories” they're interested in. What bills (entities) did a given senator (an entity-property of bills) introduce? And how are these bills' other properties similar—the subjects they affect, their passage rates, and so on? The answers tell us things about the senator who introduced them.

Adding new entities, to be used as entity-properties, expands the range of questions that analysts can answer. Did elected officials (entities) with a common donor (entity-property) vote disproportionately for a given bill (entity)? Did bills (entities) passing through a particular committee (entity-property) favor a particular agency (entity) more than other committees did?

These are rough cuts at using a formally organized schema to talk about familiar issues in public policy. Abstracting the policy process into a well organized “language” like this, then getting public policy information in machine-readable formats consistent with this language, will allow advocates, researchers, reporters, web sites, and the public to investigate questions like these, and many, many more. With structured data, the answers to public policy questions will be much easier to come by, and they will have a more solid grounding in facts and statistics than the answers we work with today.

## Two Years to a Federal Budget Data Schema

To get federal budget data in useful formats, subject matter experts should work to express the things they talk about in the abstracted language of “data” described above. Each of the things involved in the

federal budget and spending process should be reduced to description as “entities” with “properties”—with the relationships among them signaled by “entity-properties.”

The entities to be described will probably include elected officials, budget documents and line-items in budget documents, agencies and sub-agencies, government programs, bills and line-items in bills, earmarks, contracts and grants, government contractors, grantees, and so on.

Each entity will have a set of characteristics or properties that go into the entity description, with defined ways of referring to these characteristics. U.S. federal legislation, for example, has relatively standard naming conventions: a designation of the bill’s type and the house in which it was introduced followed by a number: H.R. 1234 or S. 1020, for example. A bill has only one such bill number. Likewise, bills have as a characteristic the one, and only one, Congress they were introduced in: presently, a number between 1 and 111. Was a bill also introduced in earlier or later Congresses? That’s a characteristic of a bill: having a predecessor or successor in a different Congress.

Bills may be referred to multiple committees. Thus they have as a characteristic a number of referrals defined by the set of congressional committee entities. Bills may have as a characteristic an unlimited number of “supporters” or “opponents” (entity-properties) outside the formal legislative process, though determining support or opposition is likely to be a subjective judgment compared to the objective signal given by legislators’ votes.

These are brief examples of the structured ways to describe parts of the U.S. federal legislative process. This kind of thinking should go into structuring all federal budget and spending processes over the next two years.

There are many details to think through so that the schema can capture all relevant information while maintaining flexibility and openness to extension so new, related uses of this information scheme can be adopted.

## **Ten Years to Transparent Federal Budget Data**

Over the coming decade, the transparency community should work with government agencies to ingratiate this kind of data-oriented reporting into government operations. Budgets, bills, agencies, and programs should be reported in a structured way, producing data that is consistent, machine-readable, and amenable to data processing by all segments of the policy community.

Indeed, it’s not just reporting. Each piece of the policy making process—the budgets, bills, votes, etc.—should *originate* as structured data, feeding directly into the information infrastructure that the transparency community creates. A budget should come out not just in paper and PDF version, but as a data set containing all the meaning that exists in the physical documents.

The many different parts of the policy process can adopt data-oriented reporting and origination at their own pace. As more of them do, the pieces can be woven together if an overarching schema gives the field of policymaking some consistency. There is much to be done, but the work will be easier and easier as examples of structured policymaking data come into existence and flourish.

---

Asking government agencies for “transparency” is a little bit like going to a delicatessen and asking for “a sandwich.” It begs lots of questions: “What kind of bread? What kind of meat—or veggie? Lettuce, tomatoes, onions? Oil and vinegar? Mustard? Hot or cold? Maybe toasted? Chips or a pickle? For here or to go?”

The transparency community should “place our order” by building the intellectual infrastructure for federal spending transparency, making clear exactly what information we want and in what form. This work is a small investment compared to the large dividends it will pay in decades to come. It will eliminate a current impediment to transparency by telling the information originators in government what the transparency community wants.

Data-oriented reporting of government activity will make it easier for all interest groups and actors to involve themselves in public policy and to advocate their views. Abstracting public policy processes as described here will help to create an information platform for all interests to use. Ideology, popularity, and rhetoric—the stuff of old-fashioned politics—will always have its place, but the debates that collectively determine the public interest will be better if they rest on good information made available to all.

## Appendix — The Earmarkdata.org Data Schema

What follows is the earmark data schema proposed by the Earmarkdata.org project.

### Earmark Requests

An earmark request is any communication from a senator or member of Congress to a congressional committee requesting legislative provisions that set aside funds for a specific program, project, activity, institution, or location. These measures normally circumvent merit-based or competitive allocation processes and appear in spending, authorization, tax, and tariff bills.

An earmark request is uniquely identified by its **earmarkrequestid** taken together with its **fiscalyear**.

The properties of an earmark request entity are:

- **earmarkrequestid** (*required*) A monotonically-increasing number unique in a given fiscal year. This number must be assigned by a central authority. A value of “president” is reserved and not legal in earmark request entities. (This is discussed further under Earmark entities.)
- **fiscalyear** (*required*) The fiscal year for which the allocation of funds is requested.
- **projectname** (*required*) The name of the project on which the funds are requested to be spent.
- **amount** (*required*) The amount of money requested.
- **description** A description of the purpose of the earmark requested.
- **date** The date of authorship of the request letter.
- **source** If available, a URI identifying an internet-accessible source for the information contained in the properties of this entity. Ideally, this should be the complete text of the letter.
- **[earmarkrequester]** (*required*) The entity who wrote the request letter (see below). This is always a sitting senator or representative.
- **[beneficiary]** (*one or many required*) One or more recipient beneficiaries (see below). A beneficiary is the entity for whom the funds are allocated.

### Earmarks

An earmark is a legislative provision that sets aside funds for a specific program, project, activity, institution, or location. Earmarks may be included in appropriations or authorizations bills. (*See citation below*)

An earmark request does not always become an earmark, but all earmarks should be associated with an earmark request. If an earmark is not associated with a request, that will be either an undisclosed earmark or a presidentially requested earmark. An earmark will have many properties similar to an earmark request, but these properties are not redundant since they belong to an earmark *as included in legislation*, rather than *as requested*.

The properties of an earmark:

- **earmarkid** (*required*) A monotonically-increasing number unique in a given fiscal year. This number must be assigned by the data manager.
- **fiscalyear** (*required*) The fiscal year for which the earmark directs allocation of funds.
- **type** (*required*) The type of allocation: either “appropriation” or “authorization”.
- **[earmarkrequest]** (*zero or one or many*) If the earmark originated from one or more earmark requests, these must be identified by the value of their earmarkrequestids. The referenced

earmark request entity must have been issued in the same fiscal year as the earmark entity. Where this property is present but empty, this earmark is understood to be an **undisclosed earmark**. Undisclosed earmarks are provisions that meet the definition of an earmark but that aren't expressly disclosed in report language or legislation as an earmark. If an earmark was requested by the president, the reserved earmarkrequestid "president" must be used; in this case the value does not indicate a particular Earmark Request entity but merely asserts that the president was among those who requested the earmark.

- **norequesterreason** (*required if earmarkrequestid has zero values*) If the number of earmark requests named in the earmark is zero, this property *should* be present and explain why no requesters are named or any circumstances surrounding an undisclosed earmark.
- **projectname** (*required*) The project as named in the earmark.
- **amount** (*required*) Amount of the earmark.
- **amountafteradjustment** This amount is the net amount after any congressionally-mandated adjustments such as across the board reductions.
- **description** A description of the purpose of the earmark.
- **[citation]** (*one and only one required*) An entity describing the legislation containing the earmark and its location. An allocation of funds for the same purpose or project in a different piece of legislation is a different earmark.
- **[beneficiary]** (*one or more required*) The organization receiving the earmark (see below).

## Citation

A citation is an entity describing in an unambiguous way the legislation or legislative document containing the earmark. This entity must contain all information necessary for a member of the public to locate the document and the part of the document that includes the earmark. As far as possible, this entity should have format properties or include additional properties to facilitate machine-processing so that these citations can be easily identified and cross-referenced.

- **earmarkbill** (*one and only one required*) A consistent and unambiguous identification of the bill that contains the earmark, e.g. an appropriations bill. Suggested is congressional session, followed by Thomas bill database [bill type code](#), followed by bill number, e.g. 111-H-2997. This property is to aid machine processing and cross-referencing earmark allocations with bills.
- **earmarkreport** This property is required if the earmark is contained in a report accompanying the bill. The property should a consistent and unambiguous identification of the report that contains the earmark (e.g. H. Rept. 111-123, S. Rept. 110-987, Conf. Rept. 109-231, J. Rept. 111-212). This property is to aid machine processing and cross-referencing earmark allocations with bills.
- **location** Specific location of the earmark within the document indicated by **earmarkbill** or **earmarkreport**. Includes information that enables the public to locate the earmark in the document (e.g. page number, line number, section number, subsection, paragraph, etc).
- **excerpt** Relevant excerpts of the actual language creating the earmark. This field should include language incorporating by reference any committee or conference reports or other legislative documents.
- **link** (*zero or one or more*) URLs of resources specific to this earmark.

## Earmark Requesters

An earmark requester is a senator or representative who submits a particular earmark request or earmark.

Note that if a **govtrackpersonid** property is not present, there is no guarantee that the remaining properties will be sufficient to uniquely identify the requester.

Aside from the **govtrackpersonid** property, all properties should be interpreted as applicable to that person at the time indicated by the **date** property of the earmark request.

- **govtrackpersonid** An id number from the [govtrack database](#) uniquely identifying the person. It is *highly recommended* that this property be present, since it is a well known machine-processable, unique, and temporally-unambiguous identifier of a member of Congress. If this property is present, no other properties of this entity are necessary. Conversely, if this property is absent, all other properties of this entity are required.
- **type** The type of congressman (senator or representative) of this requester.
- **state** Two-letter state code identifying the state which this requester represents.
- **districtorclass** The interpretation of this property depends upon the **type** property. For members of the house of representatives, it is the district (or 0 for at-large) which the requester represents. For senators, it is their election class (1, 2, or 3).
- **firstname** The first name of the requester.
- **lastname** The last name of the requester.

## Beneficiaries

A beneficiary is an organization to receive funds as proposed in an earmark request or as directed in an earmark.

- **duns** A Dun & Bradstreet Data Universal Numbering System (DUNS) number uniquely identifying the recipient. If present, many other properties of this entity become optional because their values are inferred from the information present in the DUNS database.
- **name** (*optional if duns present*) The name of the state, locality, business, nonprofit, or other organization receiving funds as proposed in the earmark request or as directed in the earmark.
- **address** (*optional if duns present*) Street address of recipient/beneficiary.
- **city** (*optional if duns present*) Locality of the recipient/beneficiary.
- **state** (*optional if duns present*) The state, province, or territory of the recipient/beneficiary.
- **zip** (*optional if duns present*) The zip code or postal code of the recipient/beneficiary.
- **country** (*optional if duns present*) Country of the recipient/beneficiary.

If both the **duns** property and the additional optional properties are present but the entry in the DUNS database contains information which does not correspond to the values of the optional properties, this specification does not define which set of information is more authoritative.