**Draft Proposal to enhance Privacy in Big Data systems.**

**Issues:** There is no big data privacy standard to violate. Where standards exist, they are often  
local, or ad hoc and not automatically enforced or audited. An increasing proportion of identity theft results from large systems being hacked and large sets of PII and associated attributes.

**General Proposal:**   
 Create 2 XML documents that are bound to data.  One to act to contract future behavior and provide  
 provenance. The second one provides an audit trail. Store PII separately from other data.  
   
 Provenance doc that lists

* what the data can be used for
* who has had the data
* what data has been “mashed up” with the data

Log file that lists

* Who has read/touched the file
* What transformations done on the file

**Assumption:**   In general, systems need PII for:

* Record Linkage
* Networking (special case of linkage)
* Contact user (email, phone, snail mail, Facebook page)

**Therefore:**   PII can be stored separately and leave only a Hashed\_ Id on the file.  
 Un-identified are re-identified by matching an un-identified file to another file with similar  
 characteristics and associated PII attached.

Types of access:

* Commercial
  + Marketing Analytics for segmentation, micro-targeting and direct marketing
  + Risk Mitigation products
    - Fraud detection
    - Identity validation
  + People Search products
* Research
* Surveillance, Terrorism, Law Enforcement

**Commercial**

Marketing  
 -        Typically interested in classifying and segmentation (don’t explicitly need PII).  
 -        Need PII for record linkage; after linkage can use hash\_ID and store PII separately.  
 -         Graphs are special case of record linkage.  
 -         Need PII to contact end user (including to personalize, put name in ad).  
 -         Social Networks Special Case (not sure how to handle).

* + PII and contacts are the business model
  + Assume that users agree to contract to share PII

**Research**  
 -         Typically interested in classifying and segmentation (don’t explicitly need PII)  
 -         Need PII for record linkage; after linkage can use hash\_ID and store PII separately  
 -         Need IRB to release data

**Enforcement (Regulatory, Investigation, Terrorism etc.)**  
 -          Should have a log of what was done, with appropriate log of permission to view