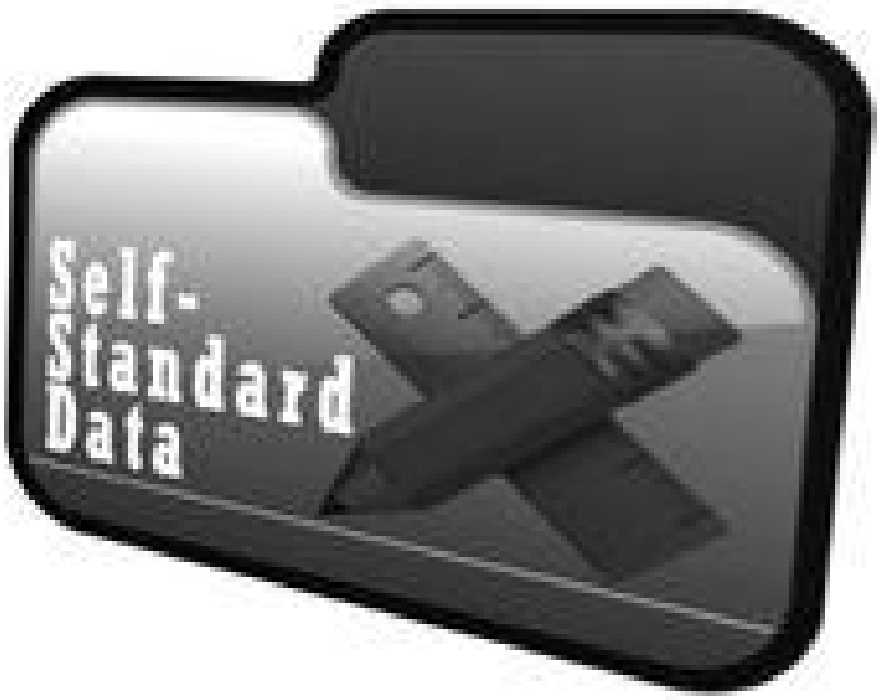
2013

自标准数据设想



2013-11-2

大庆油田 王 权

概述

自标准数据，Self-Standard Data，即自带标准或格式的数据体。数据提供者按照自己的标准或格式提供数据，并将该数据所使用的标准或格式与所提供的数据一起打包，数据使用者按照该标准或格式自行解读并使用数据。它是传统数据标准体系的重要补充和突破。

这一设想的初步想法本人于 2013 年 10 月形成。目的是为了解决“数据标准过严不易推行，过松不易整合”的问题。顺应“数据使用者急，积极性高；而提供者不急，积极性不高”的实际情况，按照客观规律充分调动数据使用者的主动性，减轻数据提供者的负担。理念一经提出即受到多位专家、学者支持，在智能数字油田开放论坛热烈讨论，在大庆油田也引起一部分人的关注，并计划在即将实施的系统中进行试验。下图是本人设计的自标准数据的标识图。

自标准数据 Self-Standard Data

定义

自标准数据，Self-Standard Data，即自带标准或格式的数据体。数据提供者按照自己的标准或格式提供数据，并将该数据所使用的标准或格式与所提供的数据一起打包，数据使用者按照该标准或格式自行解读并使用数据。它是传统数据标准体系的重要补充和突破。

性质

自标准数据打破了大家共同遵守统一标准的局限，给数据共享提供了更加切实可行的路径。其具有如下性质：

数据体，它既包含数据本身还包括数据格式；

（2）自标准数据是元数据的一种特例，元数据与数据捆绑；

（3）自标准数据是一种全新的数据共享模式，打破了传统的数据与标准脱离的局面；

（4）自标准数据是一种客观、现实的数据管理策略，适应性强；

（5）自标准数据是大数据的基本单元，采用自标准数据技术有利于大数据技术发展；

（6）自标准数据是系统自治思想的应用。

与传统数据库等的区别

自标准数据离不开传统数据库及相关技术的支持，但仍有很多明显区别：

（1）传统数据库里面的数据注重存储、查询、更新，二自标准数据更注重流动性，主要目的是数据共享；

（2）传统数据库的元数据与数据分离，查询数据时在数据库系统上分析元数据，而自标准数据中的标准就相当于元数据，它不固定在数据库上，而是随着数据体一起流动；

（3）传统数据库存储的数据量是积累性的，会越来越大，而自标准数据是增量性的，每次的体量可能变化不大，体量一般不大；

结构与数据本身加起来也可以看做是一种自标准数据体，只是体积大，不便于流动；反过来，自标准数据体可以看作是流动的数据库，只是体量较小；

（5）传统数据库重视冗余，自标准数据不重视冗余，而重视时效性；

（6）传统数据库结构是严格统一的，而自标准数据的格式和标准是允许自定义的；

（7）传统数据库主要支持某个（些）专门软件，而自标准数据主要支持系统间数据共享；

（8）与其他具体的大数据理论或技术相比，自标准数据主要是一种思想，可使用多种方式和技术实现。

起源

1998 年，大庆油田开展了一个项目——《勘探、开发、钻井数据一体化共享》。该项目目标是建立一个油田内部数据共享的平台。当时认识到，“数据使用者急，积极性高；而提供者不急，积极性不高。”鉴于此，为了实现项目目标，项目组决定顺势而为。本人是项目负责人，当时我提出一个想法，叫做“数据码头”，就是数据提供者把数据放在指定位置就不管了，使用者自己去取，去处理。使用者再产生的数据也放到码头上。这样就调动了使用者的主动性，也减轻了提供

项目组的认可。但后来大庆油田重组，勘探和钻井的大部分业务与油田开发分离，此项目下马。

那时还是要求提供者按统一标准把数据放到“码头”上，提供者还要处理数据，所以没有把提供者的工作量减到最小。当时，还没有XML，也没元数据，也没想到用它来描述数据。

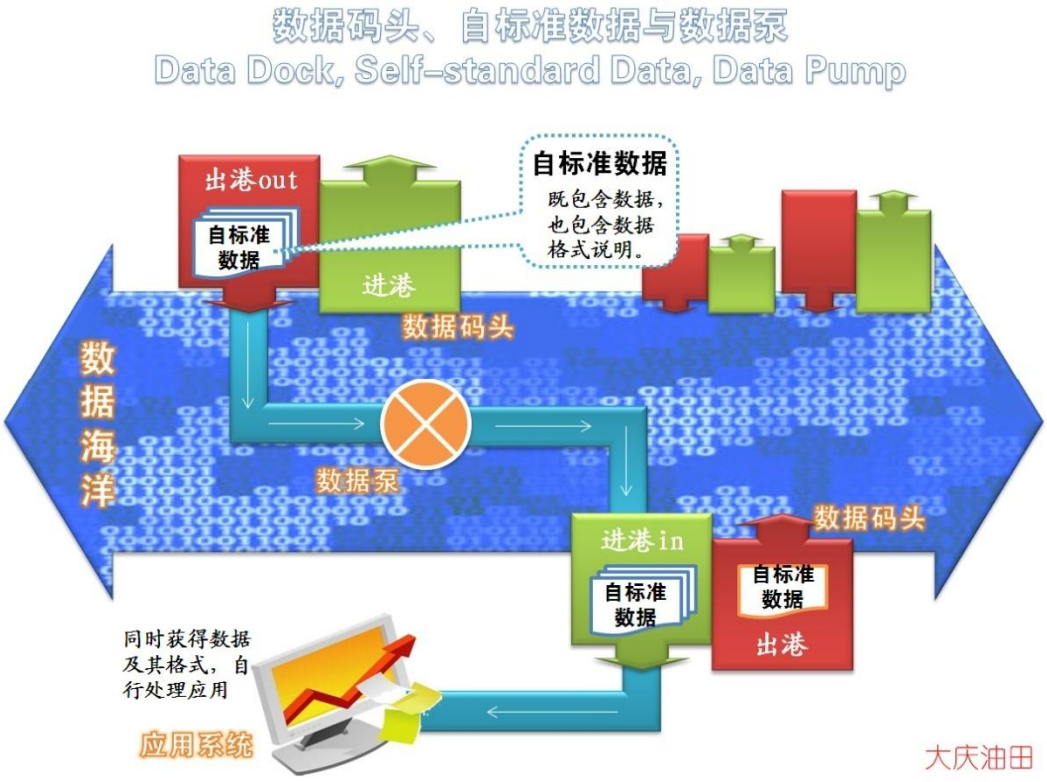
2013 年，大庆油田制定信息规划过程中，关于信息共享（十多年过去了，问题依然很多）进行了讨论。期间本人对“数据码头”思路进行了进一步的扩展，应用 XML，让提供者按照自己的格式提供数据，还要包含这些数据的格式。这样，这些数据就成为了“自标准数据”，使用者能读明白，想怎么用就怎么用。大家都方便。

2013 年 10 月，本人为长安大学数字油田论坛第三届大会准备题为《数据多了就智能!》的演讲材料时，与数字油田研究所所长高志亮教授、高倩博士进行了较深入的探讨，形成了较完整的思路。本届论坛上，各位专家学者对“自标准数据”给予了充分肯定。参加会议的多位专家学者建议，简化传统数据标准，简政放权，大力推广“自标准数据”，并认为“自标准数据”将成为大数据时代的有力的信息共享的支撑性技术。

配合自标准数据，本人进一步丰富了原来数据码头等概念，并进一步提出了数据泵的概念。

数据码头，Data Dock，即数据提供者放数据的场所。数据提供者把数据放在指定位置就不管了，使用者自己去取，去处理。使用者再产生的数据也放到码头上。

数据泵，Data Pump，是专门的抽取数据的部件，可以是集中的，或分散的。它可以被看作是传统数据适配器的改进。其功能是：存取数据，全局统一管理资源目录、使用权限等。



应用前景

数据标准一直让人头疼，主要是难以统一。

客观上，太严格不好执行，太松不好整合。

更重要的，是主观的，产生数据的人不积极遵守标准，自己方便就行了。

特别是在油田上，产生数据的人都是主角，信息化主要处于弱势的服务地位，要求勘探开发主营业务人员主动遵守标准，是十分困难的。

自标准数据有望使这一问题得到较好解决。

另外，大数据的迅速发展，传统的数据库、数据仓库、数据银行的技术都面临一个信息共享的问题，但都具有提供者不主动、使用者主动的特点，可以应用自标准数据提高海量信息共享的主动性和有效性。

致谢

自标准数据还是一个很初步的想法，能否可行尚有很多疑问，需要深入探索。感谢自本人抛出这个想法以来各位专家学者的支持、建议和批评。特别感谢高志亮教授、高倩博士、程国建教授、张艳国教授、谯英教授、黄放明教授、王哲博士等！希望有关专家、学者、技术人员、管理人员继续提出宝贵意见。

自标准数据

Q & A

2013-12-6

自标准数据 Q & A

10 月份以来，有关自标准数据的讨论越来越多，各位专家、学者都表达了自己的意见和建议。在此一并致谢！

大家在讨论迸发出了很多闪光的东西，对于我们正在面对的问题有很大的参考和指导价值。希望继续讨论。

今天，我将一些疑问和我的个人观点集中整理一下。以后会持续更新整理，并与大家共享，向大家请教，还请继续批评指导。

再次感谢！

Q1： 为什么叫“自标准数据”，它是标准么？

A： 主要基于三个考虑吧。一是“自描述”等术语已经被使用，为了避免混淆。二是“自格式”、“自定义”等有些随意。虽然是数据提供者自己定义的标准，但毕竟仍是标准，不应该随意变动。否则将出现共享障碍。“自标准”也可以叫做“自定义标准”、“自治标准”、“局部标准”，等等，其实叫什么都无所谓，简单明了就好。三是还是自认为仍然是标准，只是有所突破和侧重吧。“标准”这个字眼是不是显得更正经一点呢？ 别没事就换套衣服！稳重点行不？

Q2： 自标准数据与元数据、数据元的区别？

文上相当于“数据元+元数据”，但还是有很大的不同。首先，数据元和元数据的设计、存储、操作都有比较严格的规范，一般都是基于数据库或数据仓库的，特别是结构化数据。其次，数据元一般是不可分割的最小数据单元，而自标准数据在体量或规模上是多变的、不固定的，可以很大，也可以很小，甚至小到数据元的水平。比如自标准数据可大到一个地震工区的数据体，也可以小到只有一个井号的数据。第三，元数据和数据元一般存在于一个庞大的数据库或数据仓库实体中，一般不会同时传输。但自标准数据中数据与标准一般是在一起，就像电器与使用说明书。当然当你完全掌握了使用方法，你当然可以扔掉说明书，但那只是你自己的事，别人再用可能还得用说明书。所以说明书是标配，得跟电器在一起。

这个问题请大家多与东北石大袁满教授探讨，他在这方面造诣很深，我从他那里得到了很多指导和帮助。虽然他不同意“自标准数据”的提法，但他仍同意我的解决问题的思路。他也为大庆数字油田建设做出了很多贡献。

Q3： 为什么不用 **webservice** 或 **adapter**？

A： 我就用我当前正在面临的困境来回答吧。我们正在做一个系统，叫做《大庆油田生产经营管理与辅助决策系统》，简称 DQMDS。系统名字体现不出来建设内容。实际上我们是要建立以驾驶舱为主要功能的集成系统，要把已有业务系统的数据抽取出来，放到一块，展现出来，并一定程度地进一步钻取数据和操作。

第一个，美女只许看不许摸！初步调研显示，我们大概要集成近百个系统，最难办的是美丽的封闭系统。这些封闭系统一般都是掌管着人财物等关键资源的强势系统，还有最美的 ERP 美女。这些系统一般都是买来的，很多还是基于国外的大平台开发的，我们没有开发权限，不许我们摸。可是他们不给你接口，想建立美女热线？美得你！让服务商专门给你开发 webservice 之类的接口是很麻烦的，钱也花不起，时间也等不起，维护也耗不起。一般能提供数据就算很幸运了，而且你也别想人家遵守你的标准。所以只能把人家的数据导进来。

第二个，跟美女太亲密累得慌！我们的系统要从近百个系统里拿数据，如果都是 webservice 链接，一个系统不好使，我们就转不起来了。还是松点好，自如一些。Webservice 是紧耦合链接，实时连接，累人。自标准是松耦合，想起来就链接一下，不强迫，自己轻松，美女更轻松，连不上也死不了，可以用老数据啊，回忆也是很甜蜜的么。嗯，有点像 TCP 协议和 UDP 协议的区别！网络的链接和无链接。

但是，自标准数据不排斥 webservice、adapter 等技术，完全可以兼用。自标准数据的出发点是无奈之举，但现在看来这也挺舒服的。

Q4： 数据中心的主数据库用自标准数据合适么？

A： 不合适！数据中心是严格组织的数据，务必标准规范。自标准数据主要解决数据中心之外的数据交换问题，面向广大的人民群众，不是在洁净明亮的大玻璃房子里，而是可能在满身泥水的油田作

谁不让老百姓舒服，谁就别想舒服，领导不行！专家不行！帅哥不行！美女？？？行。。。。。。行么？。。。。。时间长了也不行！

Q5：自标准数据跟信息资源规划有关系么？

A：有！都是为了把数据或信息理顺，但信息资源规划侧重于一个系统内部，而自标准数据侧重系统间。一个是竖着的，一个是横着的，T 型。这是张艳国老师的说法，我很同意。关于这个，请多和高复先老师探讨吧，也可以和张艳国老师、胡德平先生、黄放明老师讨论，他们都是黄老师的嫡传弟子，我自命是高老师的学生，呵呵，全靠一张厚脸皮，净给高老师丢面子了。

在这里向高复先老师致敬！

Q6：自标准数据怎么考虑数据的冗余、唯一、统一、准确？

A：考虑的不多。主要考虑数据集成的现实性：让集成的系统运行起来，不被被集成的系统缠绕死，或拖死。所以松耦合是关键，那就得放弃一些严格的条件了，这是一种平衡吧。

不过，弱弱的问问：

■ 冗余都是坏处么？数据中心里不冗余就行了吧？

■ 唯一就没选择了。虽说选择有时是痛苦的，可是不同的数据之间相互还是有参考意义的吧？

■ 统一就一定好么？按照系统论观点，同质无差别的统一的系统

创新的原动力。

■ 非得那么准么？我算个周长得把圆周率精确到多少位才行呢？差不多就行吧？

■ 我们的家有必要总是整整齐齐，不敢下脚的么？非得符合几个范式？轻松点不好么？太干净的太太是不有点烦人？动不动就不让上床。同意的举手！

（随时继续添加更新）