

# 大数据出版与大数据图书馆

同方知网

王明亮

2013. 7. 11, 敦煌

# 大数据出版与大数据图书馆

- 1、一个可称之为“大数据出版”的例子
- 2、“大数据出版”有普遍意义吗？
- 3、可能的“大数据图书馆”

# 1、一个可称之为“大数据出版”的例子

## (1) “大数据”风暴——新时代露出端倪

怎么理解：出版物、数字图书馆是**大数据？进入大数据时代？**

## (2) 一个例子——中国的统计年鉴（统计数据）

《中国统计年鉴数据库》，统计出版社提供数据，CNKI出版：

1949年—2013年中国正式出版的全部统计年鉴、年报

+2010年—今 国家统计局发布的月、季度数据,反映全国各地经济结构、核算,人口、就业,投资、贸易、能源、价格、人民生活、城市、资源与环境,工农业、科教文...

共897种, 综合: 395种, 经济: 183种, 社会与文化类: 319种;

- 基本统计指标: 0.12亿个 (12M), 指标数据4亿条 (0.4G);

**总字节数**只约 150亿 (15T) == **0.015P (小数据! 1.5%)**

数据由政府依法统计、权威发布, 似乎很规范。

# 1、一个可称之为“大数据出版”的例子-2

- 如果加上衍生指标:

基本统计指标数据，只反映一部分统计对象的该项指标在不同年度的数据，并不全面，也不细致，更不具体列出已有明确定义的分析性指标。例如：

只列出“全国规模以上工业企业经济指标”、“全国国有和国有控股工业企业经济指标”，而不列出与其互补的“全国规模以上民营工业企业经济指标”；

只列出全国及各地工业企业总资产、总负载、所有者权益、总利润，但不列出资产负载率、净资产回报率。

由基本指标可以直接计算出来大量的“衍生指标”，估计至少为基本指标数的10倍！

**总字节数**== $0.015\text{PB} \times 10$ ==**0.15PB** （尺度接近大数据）

# 1、一个可称之为“大数据出版”的例子-3

- 如果加上评价指标：

统计年鉴不针对研究的问题，特别是评价问题，发布统计分析数据。如，只给“规模以上工业企业主营业务收入”、“国有和国有控股工业企业主营业务收入”，而不给出“民营工业企业主营业务收入”，更没有两者之比，不回答“国营大，还是民营大？”、“是国有经济主导，还是民营经济主导？”等问题。

类似：“民营企业与国有企业盈利能力谁强？强多少？”，“上市公司融资以后主营业务是否加速增长？”，“餐饮业收入到底有多少来自公款消费？”等等。

如果把回答这些问题的评价性指标发布出来，统计指标的数量可能还会增加5倍！

**总字节数**== $0.15\text{PB} \times 5$ ==**0.75PB** （达到大数据尺度）

# 1、一个可称之为“大数据出版”的例子-4

宏观统计指标是社会经济文化生活的宏观定量反映，尽管有种种宏观调控，但微观经济社会文化活动是在市场经济条件下随时、随机发生的，因而，宏观统计指标的数值的变化具有很强的不确定性，在这个意义上，宏观经济指标数据也具有大数据的动态性、随机性特征。

- 如果挖掘潜在数据关系：是否价值很大？

社会经济文化生活是相互影响、相互作用的，不同领域、地域的宏观统计指标之间可能存在某种必然的、或然的因果关系，而且这种关系越来越复杂，越来越隐性化，越来越难以发现，越来越速变，也越来越诱发研究兴趣。譬如，“菜价涨幅和房价涨幅有何关系？”，“期刊出版体制改革成功率是如何影响学术期刊价格的？”，“企业职工的流动率和高校专业设置的变化率是什么关系”，“CPI与城镇化率是否有关？”

# 1、一个可称之为“大数据出版”的例子-5

很多问题是想象不到的！它可能都隐含在这巨量的统计数据之中。

统计数据就像一个神秘的科幻空间，从中必可挖掘出许许多多极有价值的知识，帮助人们认识、把握千变万化的复杂社会系统。

结论：统计数据应该是 PB级以上、统计对象多样化、指标数据随机变化、潜在研究与应用价值十分巨大的**大数据**。

# 1、一个可称之为“大数据出版”的例子-6

## (3) 如何从“小数据出版”变为“大数据出版”？

- 出版观念的转变：

出版图书  出版数据（数据及其关系是更重要内容）

发布信息  回答问题（出版者做二次研究）

提供事实  服务研究（为深入研究提供工具）


- 出版手段的转变：

印刷出版  数字出版（业态转型）

表格编辑  数据编辑（数据设计，自定义指标）

宏观年鉴  微观年鉴（细化市场和产品）

数据库  研究平台（支持深度利用、价值开发）

网站大数据  网络大数据（产品与服务增值、充分开发知识服务市场）



## 2、“大数据出版”有普遍意义吗？

(1) 小文献  小数据（可传播独立内容扩容到GB级）

一本文献  碎片化  独立内容的数据化 

知识元  知识元重组  知识网络型知识块（产品）

事实性知识元：人物、事件、政策、法律、实验数据、  
情报、资源、环境、事实间关系…

理论性知识元：概念、观点、假设、原理、定律、公式、  
模型、理念…

技术性知识元：方法、技能、标准、规范、规则、程序、  
解决方案、产品、系统…

1本书的知识块数量：1  K-M，字节数：GB级。

## 2、“大数据出版”有普遍意义吗？-2

(2) 海量文献数据库 → PB级大数据

2亿篇文献 → 2000亿个知识块 → PB字节

(3) 网站PB级大数据 → ZB级网络大数据

与互联网大数据进行知识元链接，可OA，大众化。

(4) 大数据的数据关系挖掘

- 关系类型：事实与事实、事实与技术、事实与理论、理论与理论、理论与技术、技术与技术、事实+技术+理论…
- 关系属性：概念逻辑、事实因果、知识集成、学科融合…

结论：大数据知网书、面向问题的解决方案、情报服务、知识服务：将成为大数据出版的主流模式。

### 3、可能会出现“大数据图书馆”

- (1) 上游引导（出版产业转型升级驱动大数据出版）
- (2) 需求驱动（大数据科研、管理决策、学习模式）
- (3) 自身发展（服务升级、战略转型）

文献服务 → 信息服务 → 情报服务 → 知识服务 →  
研究与学习模式服务

#### (4) 大数据图书馆的特征

- 价值取向：映射真实世界，折射客观规律，为大数据条件下的创新、学习、决策活动服务；
- 资源类型、边界：多样化、多媒体、开放性大数据信息资源、大数据知网书，边界不确定性；
- 存在形态：云数字图书馆。
- 服务功能：数据搜索、重组，数据关系发现，研究平台…

## 小结

- 大数据时代是数字化时代的必然结果，认识和利用大数据的必要性、重要性已经不可置疑；
- 从大数据角度重新认识出版和数字图书馆的价值、工作理念与方式很有意义；有必要明确大数据出版的理念、概念、模式、手段，重新定位出版和图书馆的发展目标；
- 大数据出版将可能使出版产业进入大数据市场，重新激活数字出版，促成出版产业的数字化转型；
- 大数据图书馆将应运、应势而生，并有可能很快成为现实。

谢谢大家！