Analysing data is the future for journalists, says Tim Berners-Lee
Inventor of the world wide web says reporters should be hunting for stories in datasets

Sir Tim Berners-Lee reckons he's glimpsed the future of journalism – and given he's the person who invented the world wide web, you might not want to bet against him.

In his view, it lies with journalists who know their CSV from their RDF, can throw together some quick MySQL queries for a PHP or Python output ... and discover the story lurking in datasets released by governments, local authorities, agencies, or any combination of them – even across national borders.

That's because he thinks the future lies in analysing data. Lots of data. Speaking on Friday at the launch of the first government datasets for spending by departments of more than £25,000, he was asked who will analyse them once the geeks have moved on. What's the point? Who's really going to hold government, or anyone else, accountable?
"The responsibility needs to be with the press," Berners-Lee responded firmly. "Journalists need to be data-savvy. It used to be that you would get stories by chatting to people in bars, and it still might be that you'll do it that way some times.

"But now it's also going to be about poring over data and equipping yourself with the tools to analyse it and picking out what's interesting. And keeping it in perspective, helping people out by really seeing where it all fits together, and what's going on in the country."

If that sounds like a daunting prospect, then it's worth considering that hardly any of the journalism courses today teach any sort of data analysis – not even its simplest form, statistics.

But that might be changing. Earlier this month City University launched its MA in interactive journalism, led by Jonathan Hewett and Paul Bradshaw, which will teach "data journalism" as part of its curriculum – "sourcing, reporting and presenting stories through data-driven journalism, and visualising and presenting data (including databases, mapping and other interactive graphics)."

Bradshaw says there are elements in courses at Birmingham, Cardiff and Southampton. "There's an awareness of it, but there are so few people even in news organisations who can do it. People are frantically trying to catch up and get their heads around it."

It's easy to see how Berners-Lee could be correct in his forecast both about what journalists do, and how they could benefit. Think of the biggest story of the past two years, on MPs' expenses. That grew out of Freedom of Information requests posed by Heather Brooke, which then led to the Daily Telegraph's bought-up exclusive involving a team of people poring over data, followed by a public release of MPs' receipts in which the Guardian, among others, recruited the public – with the help of custom-built tools – to try to analyse what had gone on.

And of course Tony Blair, in his memoirs, recollects his belief that FOI was misused: "Freedom of information. Three harmless words. I look at those words as I write them, and feel like shaking my head till it drops off my shoulders. You idiot. You naive, foolish, irresponsible nincompoop. There is really no description of stupidity, no matter how vivid, that is adequate. I quake at the imbecility of it ..."

"The truth is that the FOI Act isn't used, for the most part, by 'the people'. It's used by journalists. For political leaders, it's like saying to someone who is hitting you over the head with a stick, 'Hey, try this instead,' and handing them a mallet.

"But another and much more important reason why it is a dangerous act is that governments, like any other organisations, need to be able to debate, discuss and decide
issues with a reasonable level of confidentiality. ... Without the confidentiality, people are inhibited and the consideration of options is limited in a way that isn't conducive to good decision-making."

Francis Maude acknowledged on Friday morning that releasing the data about spending over £25,000 could be an act of masochism – a suspicion that much of the early coverage confirmed.

But it is probably only the beginning – and it is likely that journalists won't be the first who really dig into the data with most effect. Although the Guardian, Telegraph and Times all have data teams who aim to find stories in big datasets, such as the Guardian's geotagged coverage of the Wikileaks documents from Afghanistan and Iraq, or the Telegraph's analysis of the London Bike Hire scheme, "Most of the innovation is happening outside news organisations," Bradshaw says. "Sites like Openly Local, Charities Direct, Who's Lobbying?, Where Does My Money Go? and Scraperwiki. They're all hiding their light under a bushel. All doing great things."

But how long will it take for the methods of data journalism – where CSV (comma-separated value files, a form that any database or spreadsheet program) and RDF (Resource Description Framework, a way of linking different data sets) and MySQL (a free, open source database program able to cope with tiny or huge datasets) and PHP (a programming language widely used to write web pages) and Python (another web programming language) are part of the landscape – to filter through to everyday use in journalism? As William Gibson observed of the future, it's here already, just not very evenly distributed. Bradshaw says that the Press Association is "definitely interested" and magazine publishers also want to adopt data journalism techniques.

Possibly there will be more detail lurking in the expenditure published last week, or in future versions, which Maude says will become more detailed. It's clear there's going to be plenty to study.

Yet will it obviate the brown envelope in the smoky bar – the place where Andrew Gilligan got his tipoff from David Kelly about No.10's dodgy dossier, where perhaps the Telegraph got the MPs' expenses hard drive, where countless other stories have been whispered? Perhaps not.

"Data-driven journalism is the future," Berners-Lee insisted. To which his colleague Nigel Shadbolt, who with Berners-Lee has been working to get the civil service and local government to open up their data, added succinctly: "Well, part of the future."