Forecasting Terrorism:
Meeting the Scaling Requirements

Steven R. Newcomb (srn@coolheads.com)

Extreme Markup Languages 2002

August, 2002, Montréal

Executive Summary

The internationally standardized "Topic Maps" and "GROVE" paradigms offer unique advantages to the detection of terrorist threats. These paradigms allowthe knowledge output of existing processes to be integrated, at any scale and at any granularity, without first having to modify either the processes or their outputs. The paradigms suggest ways to eliminate certain kinds of bottlenecks, allowing governments to shorten the time required to integrate new knowledge by increasing the number of computers working on the problem.

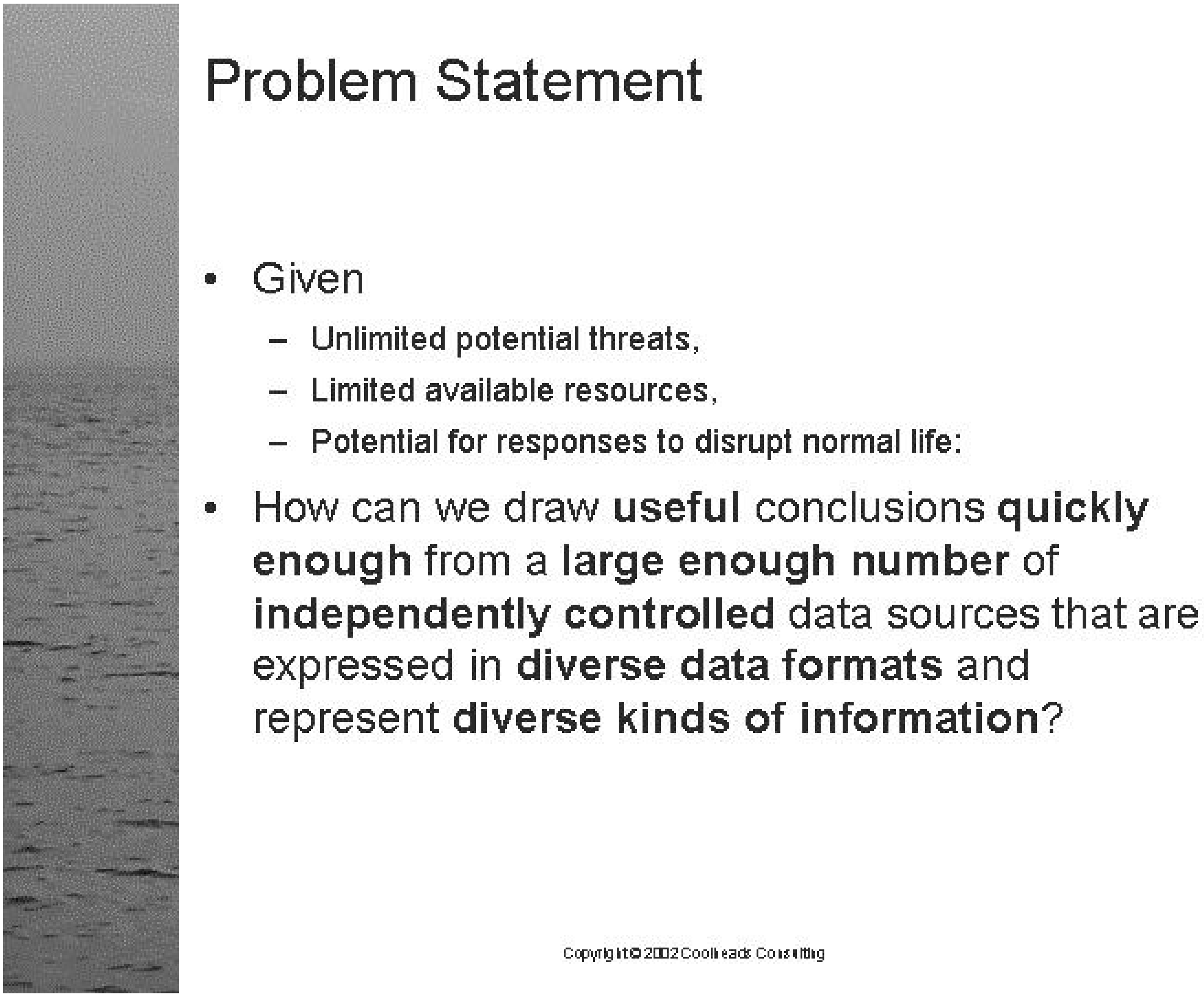
(Steven R. Newcomb serves as editor of several international standards for information management, including ISO/IEC 10744:1992 and :1997 ("HyTime" Hypermedia/Time-based Structuring Language), and ISO/IEC 13250:2000 ("Topic Maps"). He is a founding co-chair of the Extreme Markup Languages annual conference series hosted by IDEAlliance, and he works as a consultant with Coolheads Consulting.)

Introduction

- “No” to terror, “Yes” to freedom
- “Who will guard the Guardians?” -- Plato
- It’s an information management problem.

It would be a good thing if terrorist attacks could be avoided, without undue expense, and without the necessity of resorting to war, colonialism, the curtailment of liberty, or the destruction or distortion of the intricate and delicate webs of relationships required to support free trade and open societies.

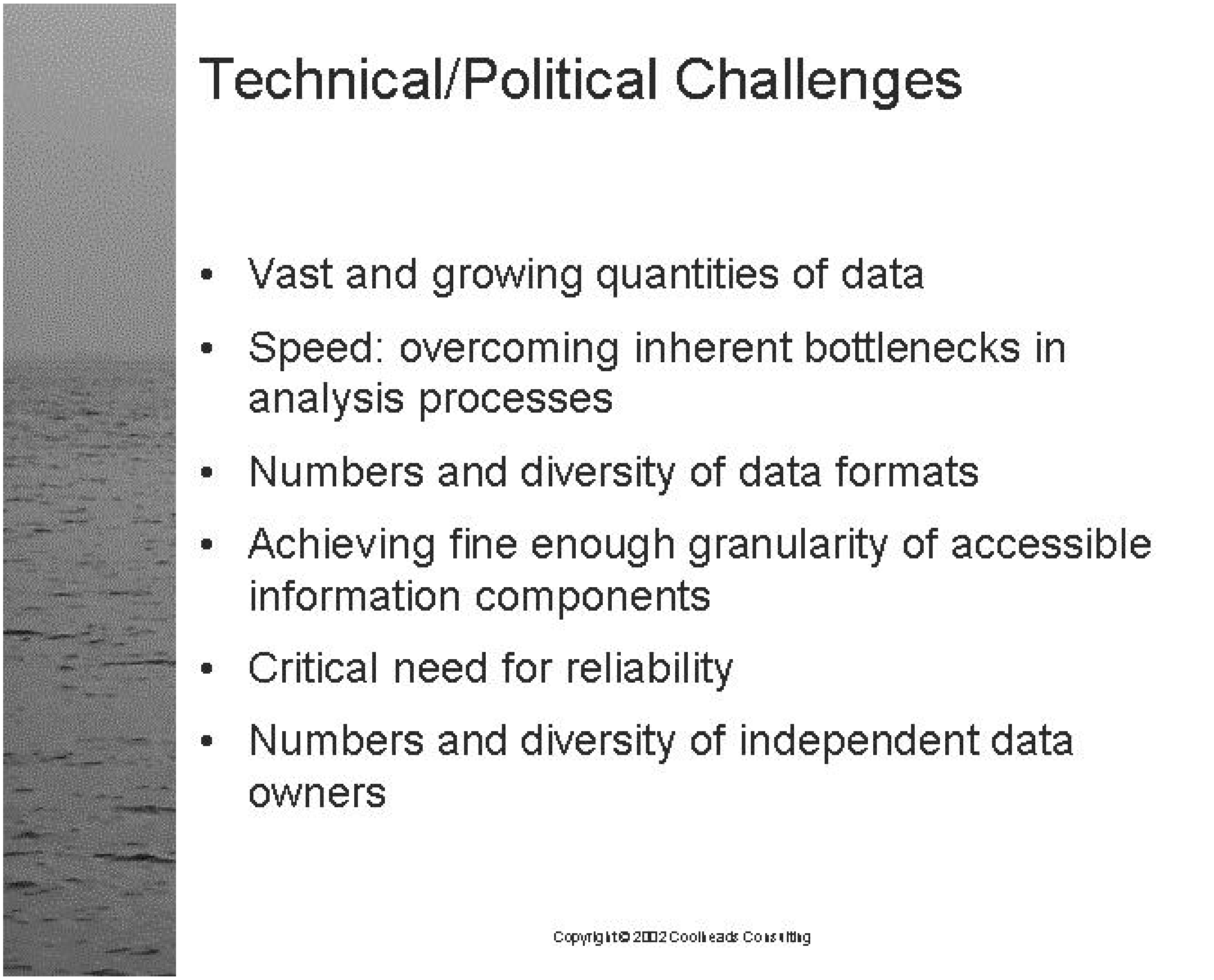
It's a thorny information management problem with extreme technical requirements, and the adaptations that will need to be made in order to cope with the social implications of meeting the requirements are, if anything, even more extreme. It would be fascinating, for example, to consider the information management problem from the perspective of avoiding a situation in which too much power (in the form of economically significant knowledge) is concentrated in the hands of too few people. Science fiction has long speculated about the structures of societies in which the idea of privacy has been redefined in various provocative ways. The 2,500-year-old question posed in Plato's *Republic*, "Who will guard the Guardians?" is at least equally relevant. However, the question of what the technical requirements are, and how they can be met, is slightly more urgent. Virtually all of the information that supports civilization may need to be integrated before the tenuous patterns of terrorist behavior can be reliably and comprehensively detected.



Problem Statement

- Given
  - Unlimited potential threats,
  - Limited available resources,
  - Potential for responses to disrupt normal life:
- How can we draw **useful** conclusions **quickly enough** from a **large enough number** of **independently controlled** data sources that are expressed in **diverse data formats** and represent **diverse kinds of information**?

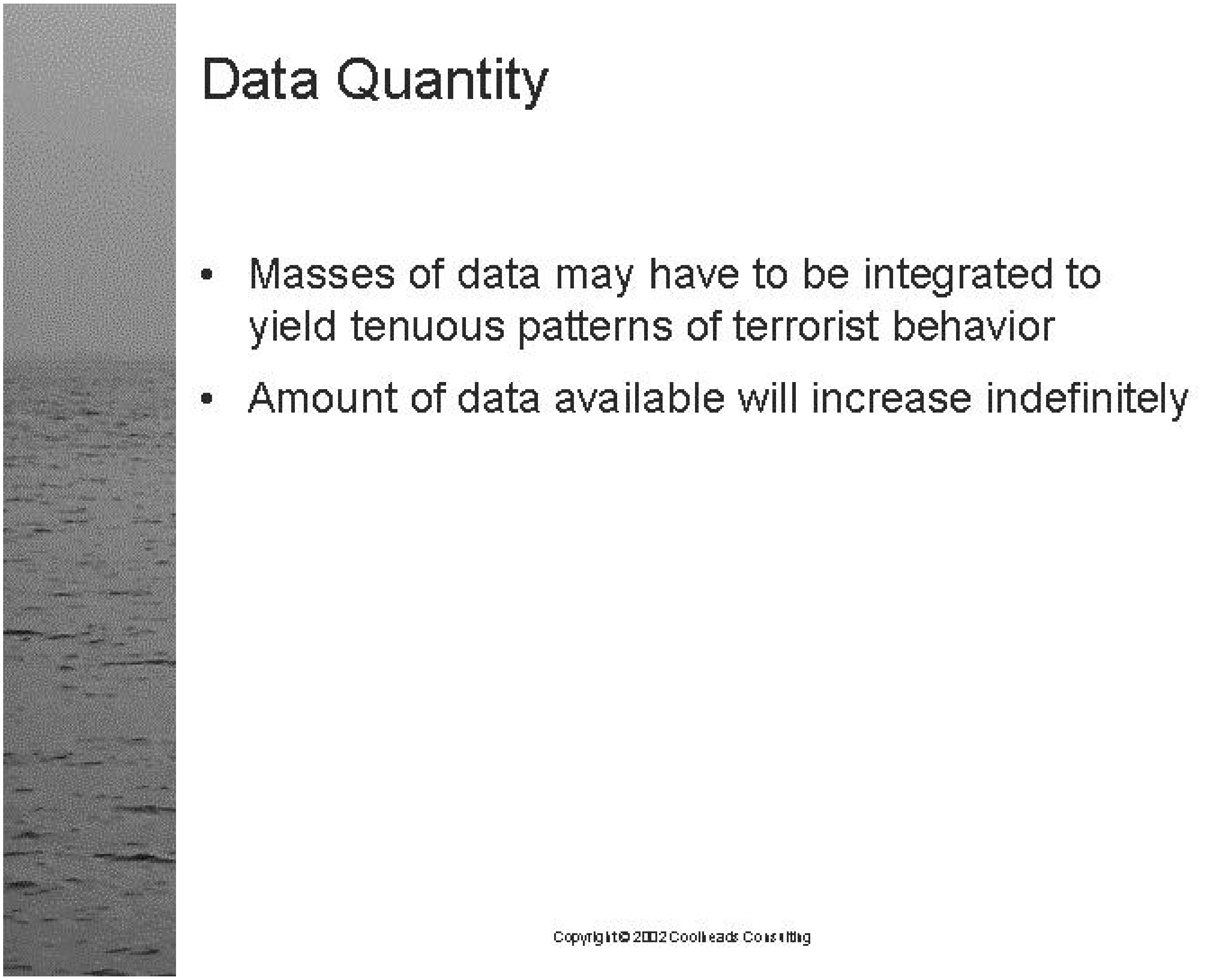
When a possible terrorist threat has been detected, and when a decision ismade to try to avert an attack, considerable expense may be incurred, and the normal activities of many people can be disrupted. Moreover, the resources that can be used to respond are limited, while there is no limit on the number of possible threats that can arise simultaneously. In other words, the decision to respond to a possible threat may be very consequential. All ofthe data that ledto the detection of the threat must be rapidly verifiable, and the reliability of each datum must be re-assessable. From the perspective of any conclusion about the existence of a threat, then, it is vital to be able to "see" the sources on which the conclusion was based. Conclusions based on evidence that cannot be verified, and that cannot be fully re-assessed, are not useful.



Technical/Political Challenges

- Vast and growing quantities of data
- Speed: overcoming inherent bottlenecks in analysis processes
- Numbers and diversity of data formats
- Achieving fine enough granularity of accessible information components
- Critical need for reliability
- Numbers and diversity of independent data owners

Let's consider these challenges in terms of the technical requirements each of them implies.

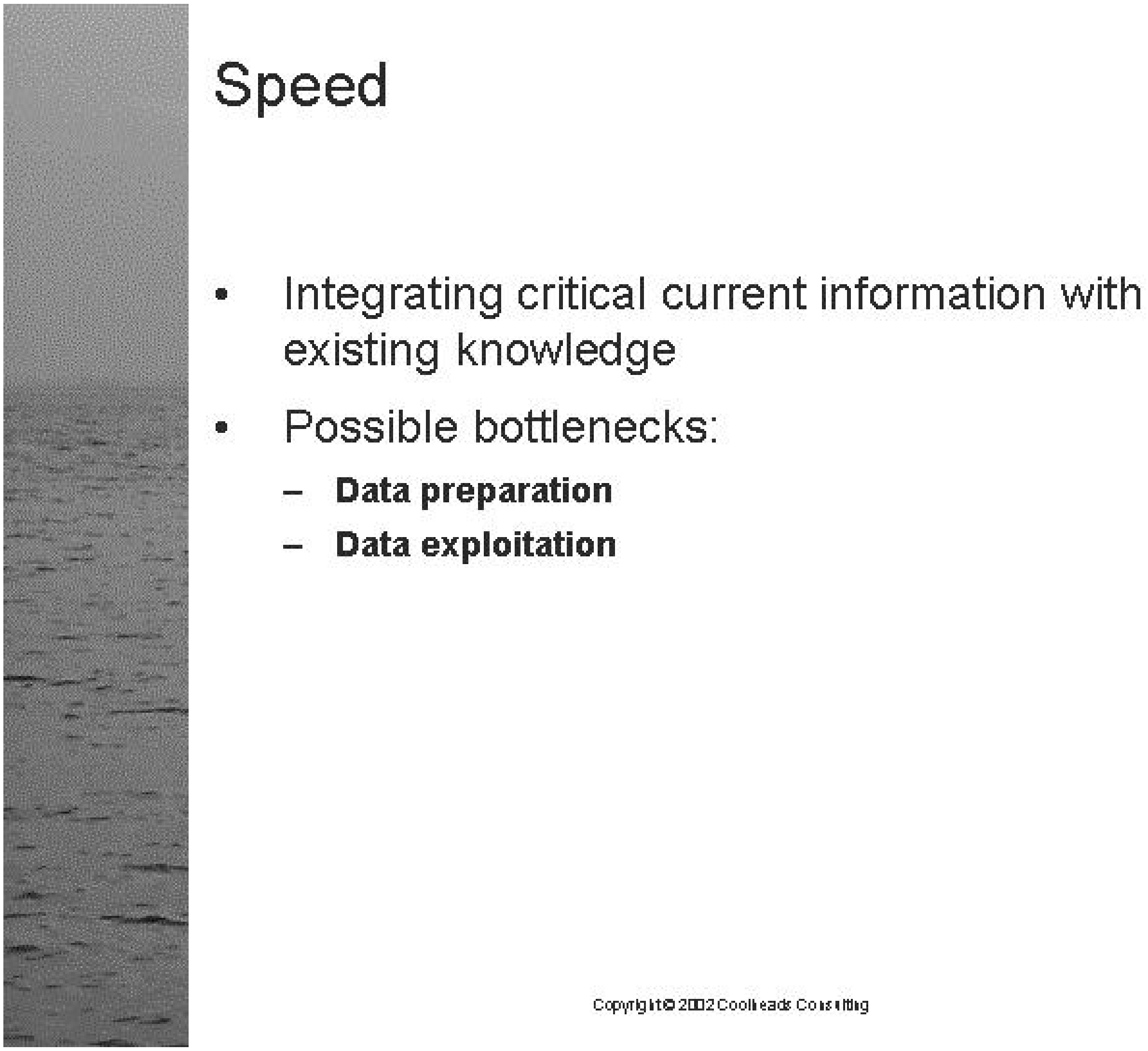


Data Quantity

- Masses of data may have to be integrated to yield tenuous patterns of terrorist behavior
- Amount of data available will increase indefinitely

The quantities of information that must be integrated can be measured only inastronomical numbers.

No assumptions about limits on the quantity of information to be integrated are big enough to be safe. The amount of information to be integrated is certainto increase indefinitely. The rate at which new information is created will also increase indefinitely. As our activities are monitored in ever-increasing detail, there will be occasions whenentirely new avalanches of information begin to become available. Some ofthese new avalanches may suddenly and very greatly increase the overall daily amount of information that must be integrated.



Speed

- Integrating critical current information with existing knowledge
- Possible bottlenecks:
  - **Data preparation**
  - **Data exploitation**

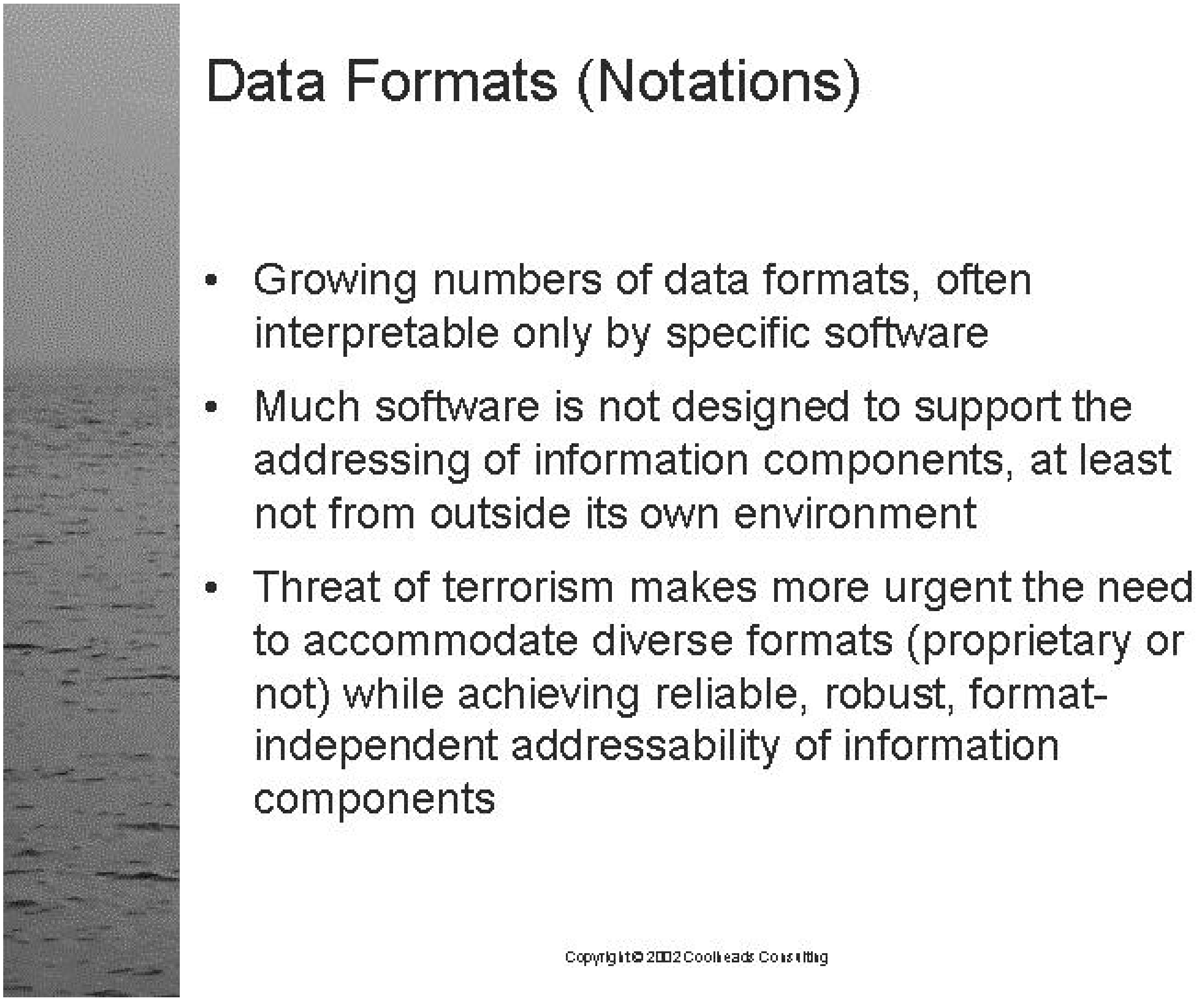
Critical information may become available onlyjust before or during aterrorist attack. The speed with which new information can be understood interms of what is already known, thereby to detect and avert a disaster, is critical. There is not only the usual budgetary concern about the number of expended human-hours of effort, but also a tactical consideration: the number of minutes of elapsed time required to integrate new information and understand its implications in terms of terrorist threats.

There are two distinct time-critical operations:

1. **"Data preparation:"** Pre-processing operations that allow intelligence to be drawn from the resulting data in a fashion that is consistent with allthe other constraints and requirements. Data preparation time is limited; knowledge must be able to be integrated without first having to copy, convert, and warehouse it. Society's security interests are best served by using the most current version of any information asset. For many kinds of relevant information assets, the time required to convert and warehouse them guarantees that the converted, warehoused form does not reflect the most current version.

2. **"Data exploitation"** means drawing conclusions from the prepared data.

It's a requirement that both of these operations be accomplished within alimited human-hour budget and within an extremely limited period of elapsed time.



Data Formats (Notations)

- Growing numbers of data formats, often interpretable only by specific software
- Much software is not designed to support the addressing of information components, at least not from outside its own environment
- Threat of terrorism makes more urgent the need to accommodate diverse formats (proprietary or not) while achieving reliable, robust, format-independent addressability of information components

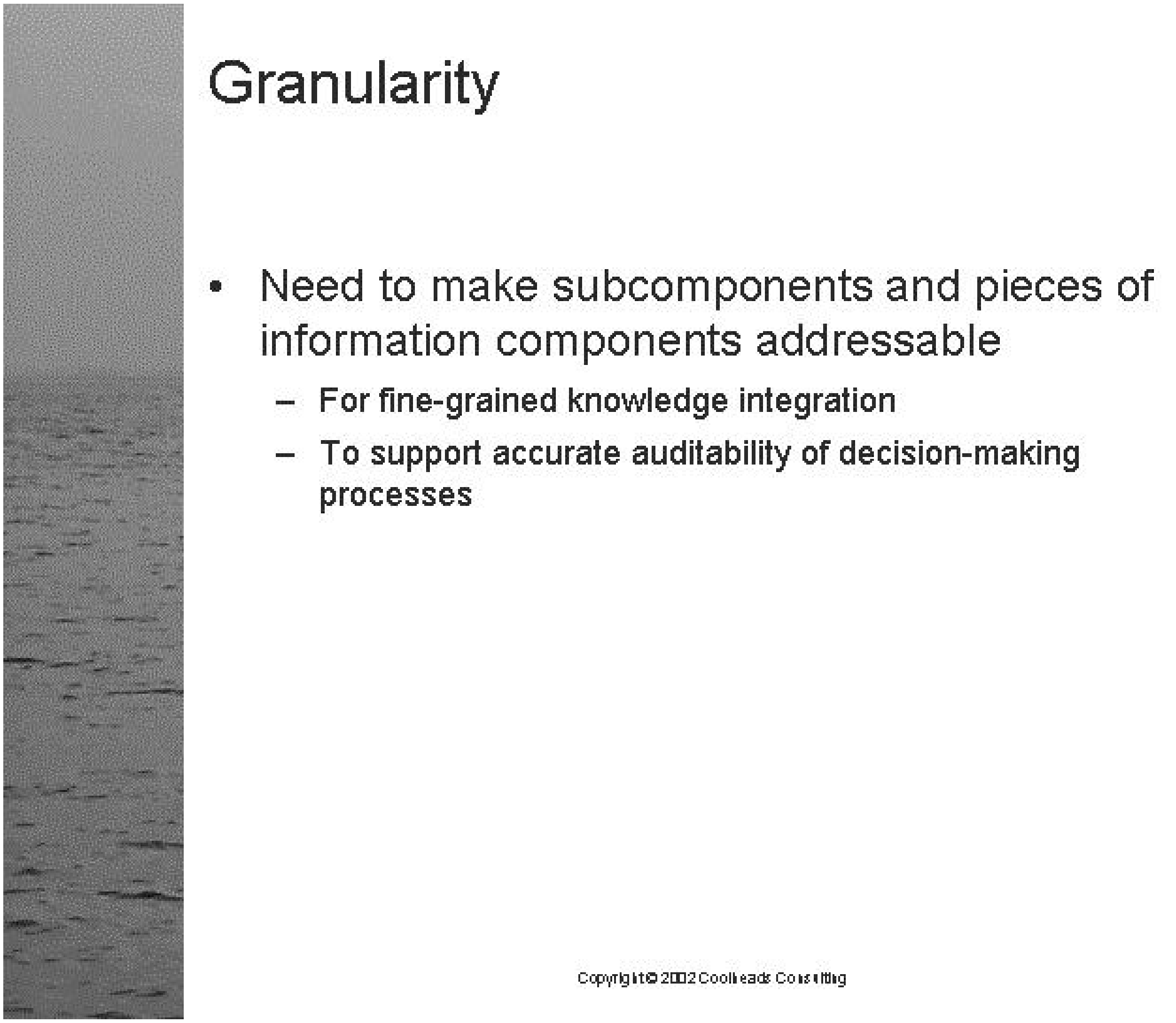
Most data is interpretable only by specific software. Much software is not designed to support the addressing of information components. When proprietary software does provide for component addressing by other applications, it generally does so only within an operating environment that is contrived by the vendor in such a way as to discourage or prevent the addressing of information components from outside that controlled environment.

Nevertheless, analysts need to be able to address any piece of information within a single apparently-unified address space. Any lesser solution would waste time and effort.

However, it is not practical for the government to impose a single addressingconvention on all information owners, and on all their systems. Either the government must incur the cost of making the information placed at its disposal addressable, or it must pay information owners to make their reports conform to the government's addressability needs. In most cases, and especially given the other requirements that must be met, it would be less expensive for the government to adapt its systems to the various evolving internal and industrialconventions of information owners, rather than the other way around.

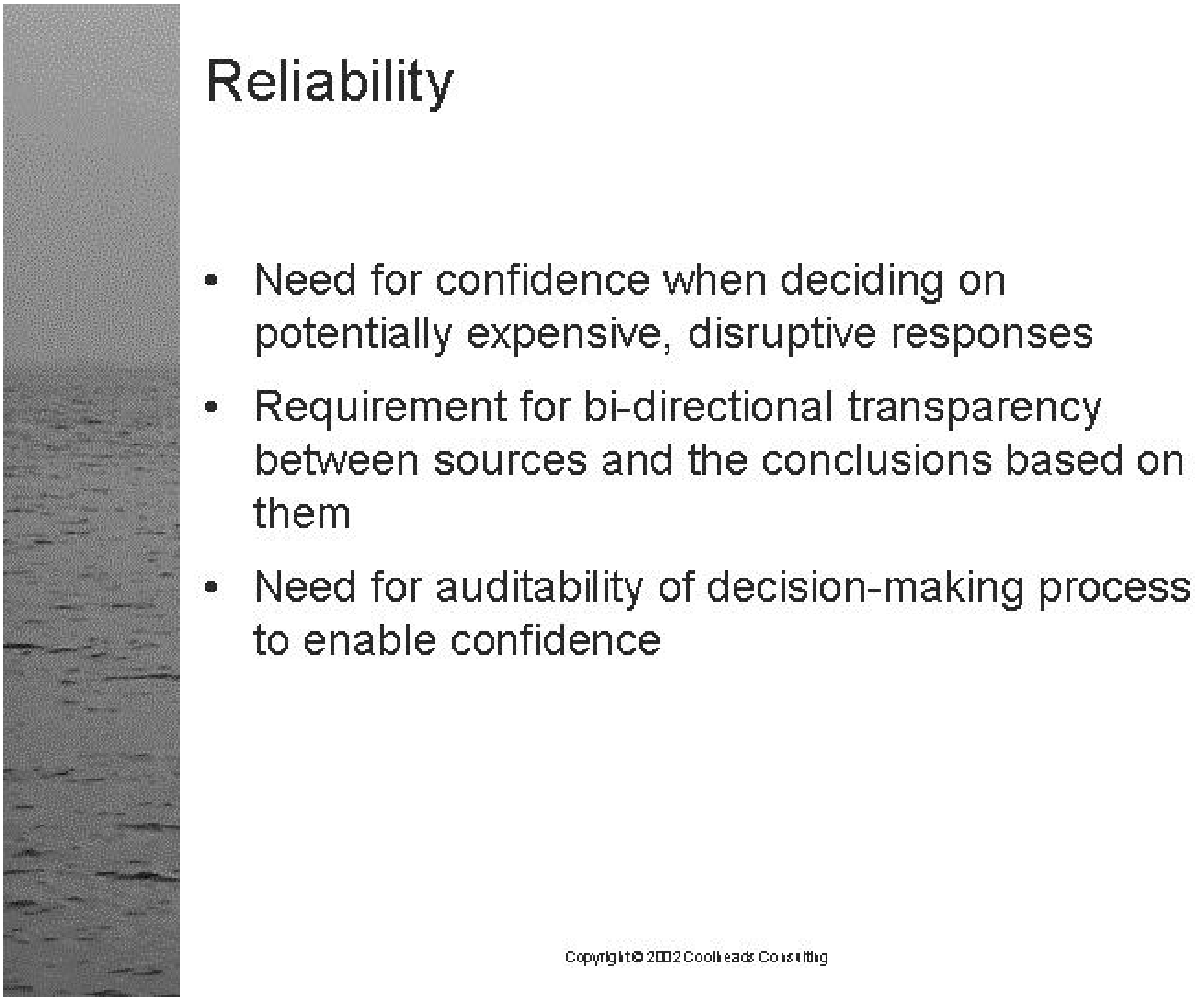
The total number of information formats (*notations*, in the parlance of SGML and XML) that the government must be able to bring under its universal addressability umbrella will continue to grow indefinitely. For example, there are about as many notations for .doc files as there are versions of Microsoft Word, and new versions of Microsoft Word will use more different notations.

The question of how an unboundedly diverse corpus of information can be addressed, and what it means for an arbitrary information component to become addressable, must be answered in light of the aforementioned requirement for transparency between evidence (the various pieces of information available to the government) and conclusions. Ideally, the information handling methodology used by governments should provide an effective link between each piece of evidence and each conclusion it putatively supports, and between each conclusion and each piece of evidence that supports it, but without changing either one, and without negatively impacting their manageability, or the freedom of information owners to conduct their businesses in any way they see fit, using any information technologies they choose to use.



Granularity

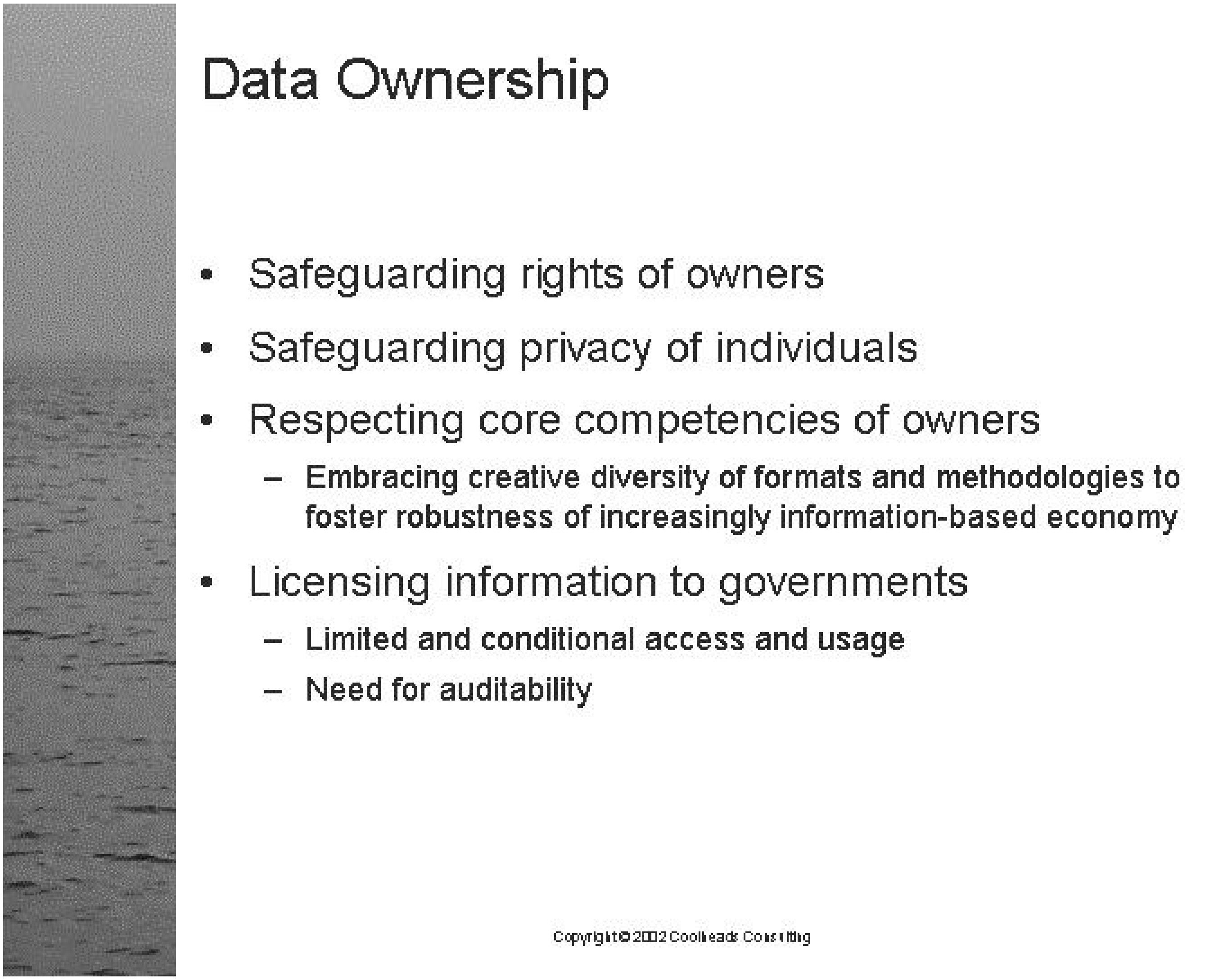
- Need to make subcomponents and pieces of information components addressable
  - For fine-grained knowledge integration
  - To support accurate auditability of decision-making processes



Reliability

- Need for confidence when deciding on potentially expensive, disruptive responses
- Requirement for bi-directional transparency between sources and the conclusions based on them
- Need for auditability of decision-making process to enable confidence

Sources of intelligence are often associated with reliability estimates. These estimates are subject to change, and, when such changes occur, the reliability of older information-- and the conclusions that such information has impacted -- can be affected. From the perspective of any source, it is vital to be able to "see" the conclusions that were affected by it. If the list of affected conclusions cannot be determined, it will be impractical to respond to changes in reliability estimates, because it is not feasible to draw all conclusions over again, every time a reliability estimate is revised for one of the sources.



Data Ownership

- Safeguarding rights of owners
- Safeguarding privacy of individuals
- Respecting core competencies of owners
  - Embracing creative diversity of formats and methodologies to foster robustness of increasingly information-based economy
- Licensing information to governments
  - Limited and conditional access and usage
  - Need for auditability

Much of the information that may be relevant to detecting terrorist threats is private property that may be made available to government agencies under special arrangements that safeguard the rights of the owners, and that are consistent with the owners' various other responsibilities, such as their responsibility to safeguard the privacy of individuals to whose lives the data are relevant. It's really just a question of supporting limited licensing of the information to the government. The cooperation of owners of interesting private information, such as educational institutions, private libraries, landowners, news organizations, industrial consortia, NGOs, etc. can be purchased by governments. Such owners would naturally impose conditions and require various kinds of assurances, and the government would have to live with various complex limitations on its ability to access and use the licensed information. The complexity of supporting a government's ability to adhere strictly to the conditions under which it receives information underlines the general requirement that information handling processes be auditable.

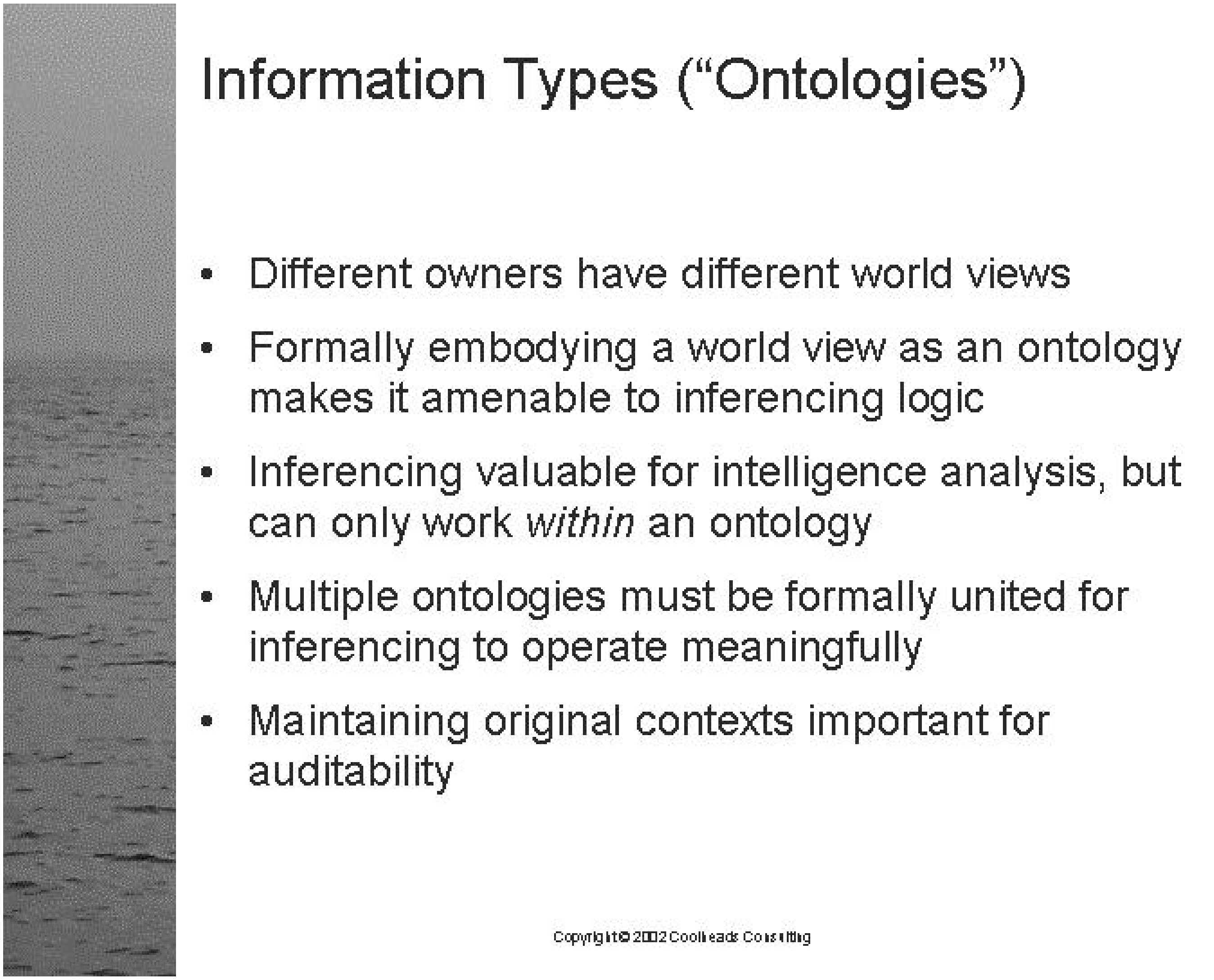
Even among the intelligence and law enforcement agencies of a single government, there are intramural sovereignty issues that must be addressed. Each different agency is itself a sovereign information owner whose concerns must be respected by the rest of the government. Today, at least within the U.S. government, there is considerable pressure on various agencies to share information with each other, but it is also true that each has different missions, responsibilities, and authorities. Each requires specific (and sometimes unique) assurances from any entity with which it shares its information, even if such entity is another government agency.

There is another aspect of the control of information owners over their information assets that should be respected by the government: the government cannot hope to constrain the data representations, asset maintenancemethodologies and storage/retrieval systems. The choices made by information asset owners in these areas reflect the core competencies of their businesses -- core competencies that the government does not have. Perhaps even more to thepoint, the imposition of uniform technological or methodological requirements could damage the long-term competitiveness of an increasingly information-based national economy.

The need to respect the independent control of information owners over their own information assets translates into the following specific technical requirements:

- A government's information handling systems should be capable of enforcing arbitrary rules regarding access and use of privatelyowned information. This enforcement must work very reliably, because when it fails, the consequences for the information owner may be dire.
- A government's information handling systems should support the auditing ofthe uses to which information has been put.
- Analysts must be able to integrate knowledge without first having to copy, convert, and warehouse it. Government warehousing of knowledge implies government possession of it, and at least some potential information supplierswill be unwilling or unable to comply with such a requirement. A government must be able to "lazily" integrate knowledge that is held in systems that it does not own or control.
- The design of a government's intelligence information handling systems cannot assume that the government will impose constraints on data representations, maintenance methodologies, or storage/retrieval systems. Thisis a challenging requirement. If the process of integrating and refining knowledge is fully auditable, it must be true that a single addressing expression language -- a "single address space" -- is in place, and is capable of addressing all of the information that plays any role in the process. This requirement comes from the fact that any conclusion may have been drawn fromany combination of evidence, housed in any repositories, and represented by any notations, syntaxes, or data structures. The semantics and capabilities of such a single uniform address space must be able to be adapted to encompass anything new that comes along without compromising any existing addressing expressions. New repositories, notations, data structures are created constantly and any of them may turn out to be relevant to the problem of detecting terrorist threats.

The latter requirement is particularly challenging, even though the standards and technologies needed to penetrate the ghetto walls embodied by specific data representations, storage systems, and ontologies already exist. (I'm referring here to the GROVE paradigm, the subject of ISO/IEC 10744:1997 Annex A.4, and the Topic Maps paradigm, the subject of ISO/IEC 13250:2000, and their implementations.) The information that our governments normally require us to report, such as tax and banking records, is required to be reported in conformance with rigid format constraints. These constraints (such as thoseembodied by tax forms) greatly increase the efficiency with which a government agency can operate. However, a world in which literally any information is regarded as potential input to one or more government agencies cannot presuppose that there is already a government agency form for each kind of reportable information. One challenge facing those who would integrate diverse knowledge is the widespread presumption that bureaucracies must always retain authority over their input formats -- over the forms to be filled out by their various officials and supplicants. This bureaucratic presumption has been a cornerstone of successful human civilizations for thousands of years. It will take time for governments, and perhaps especially for the peoples they govern, to adjust to the idea that, with comparatively few exceptions, the authority to design a government form must, in effect, be delegated to every information owner. Since the survival of civilization depends on both accurate communications and the conservation of diverse ways of doing business, government bureaucracies must prepare themselves to handle information inputs that are diverse and constantly increasing in their diversity.



Information Types ("Ontologies")

- Different owners have different world views
- Formally embodying a world view as an ontology makes it amenable to inferencing logic
- Inferencing valuable for intelligence analysis, but can only work *within* an ontology
- Multiple ontologies must be formally united for inferencing to operate meaningfully
- Maintaining original contexts important for auditability

The nature of knowledge is that it has no specific nature. Instead, it has a number of natures that is as unbounded as the potential number of human individuals, multiplied by their various moods and levels of understanding. Knowledge always has at least one context, and without context, it cannot exist.

Different individuals have different world views -- different contexts -- and they cannot understand each other except to the extent that their world views logically converge.

A "universe", "context", or "world view" can be formally embodied by an"ontology" -- a set of types of relationships between the things (sometimes called "knowledge entities") that exist in that universe. Inferencing -- the use of logic to make implicit knowledge explicit -- must be based on rules regarding how constellations of knowledge entities and the instances of relationship types that connect them together can be interpreted. Inferencing is a very valuable tool for intelligence analysts, but it can only work within an ontology. Inferencing cannot work on constellations of relationships whose types are declared in different ontologies, unless those different ontologies have been formally united in a single ontology. Such unification is analogous to creating an additional universe that at least partially encompasses all the universes in which the relationships within the constellation exist, and thus provides a context in which all the pieces of knowledge can be considered by a singleinferencing process. It is not always possible to unite different ontologies, because of fundamental incompatibilities between the world views that they reflect.

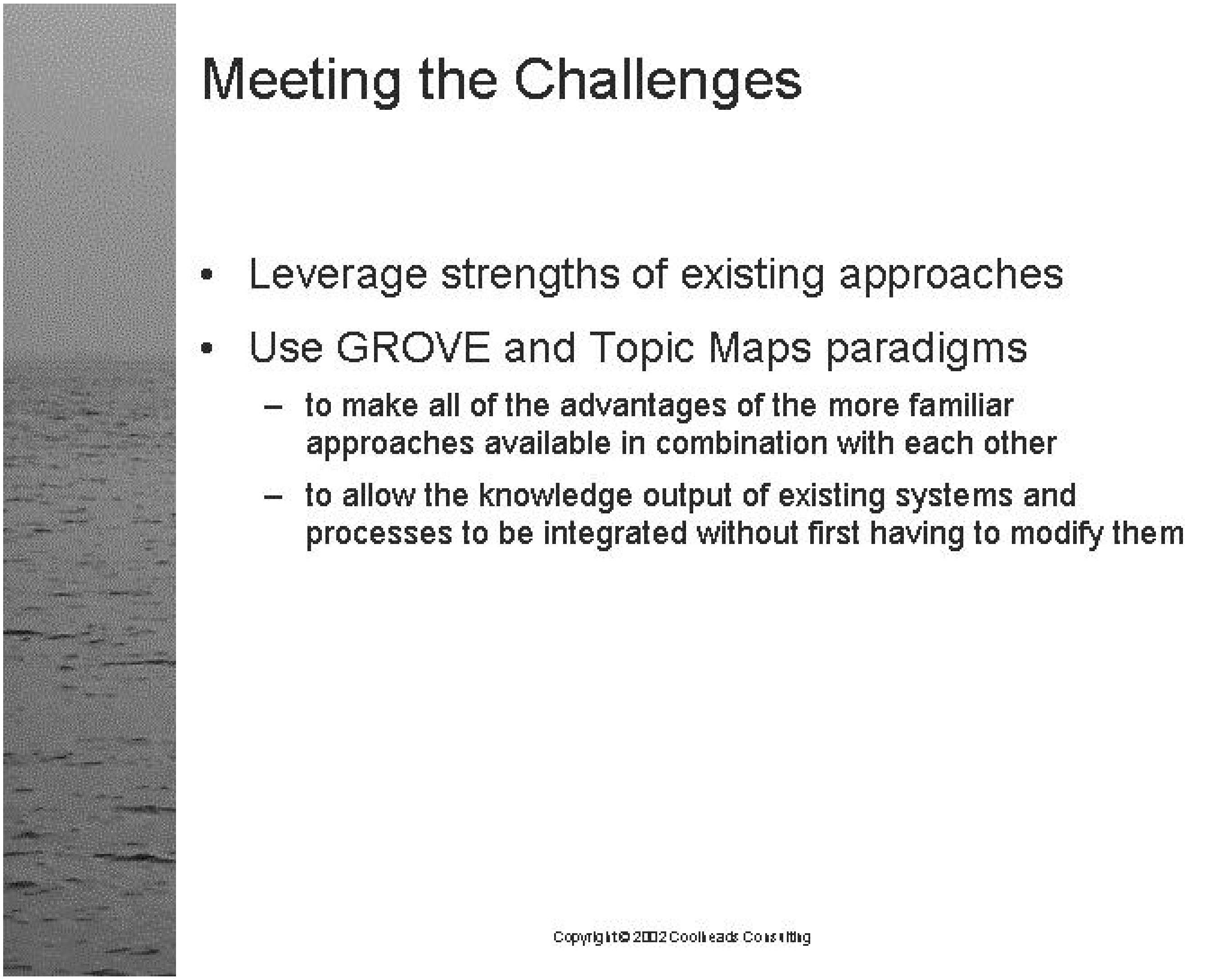
The requirement that must be met by systems intended to support anti-terrorist analysts is, first of all, that it not seek to impose a single world view or ontology on all information owners. Instead, as in the case of diverse information formats, governments must be prepared to accept diverse knowledge, expressed in the terms (i.e., the ontologies) of the diverse world views of its owners and maintainers.

Another requirement is that pieces of knowledge must never be separated from the contexts in which they must be understood. It must always be possible to determine which world view is the one within which any given statement was made. (This can be seen as yet another auditability requirement.)

In order to make knowledge emanating from any given world view as useful as possible to analysts who may not be familiar with it, it is reasonable to require that all relationships be self-describing in terms of their types, and that all relationship types be self-describing in terms of the world views they comprise.

As in the case of diverse information types, new ontologies are born frequently. It is a requirement that there be no limit on the number ofontologies that can be supported by intelligence systems.

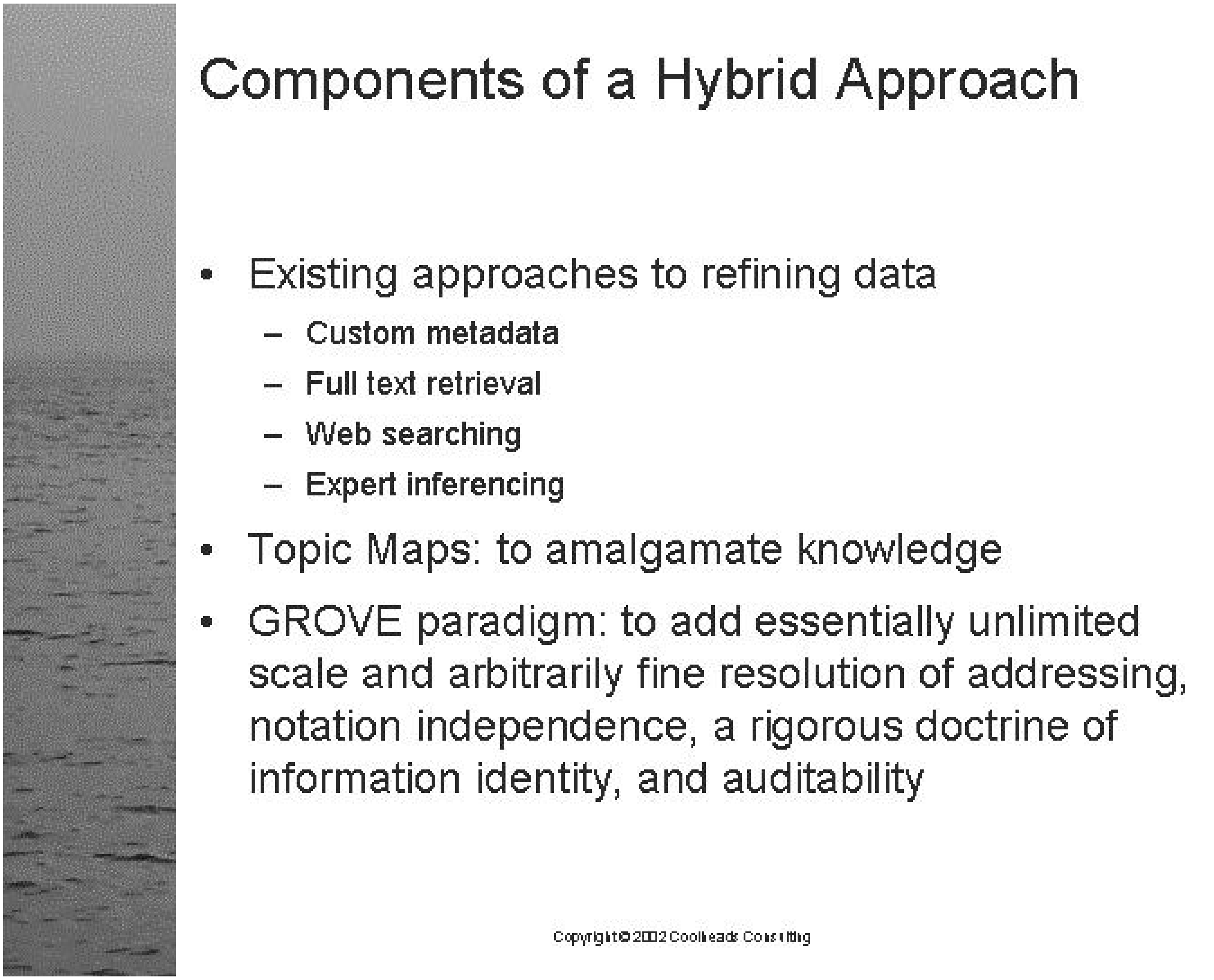
There needs to be a single address space within which all subjects are addressable. If this requirement is not met, there will be no way to allow analysts to amalgamate all knowledge, regardless of its combination ofontological contexts, that is relevant to any given subject. This single address space must encompass subjects that are themselves pieces of information, as well as subjects that are not pieces of information, and it must preserve the distinction between the two kinds of subjects.



Meeting the Challenges

- Leverage strengths of existing approaches
- Use GROVE and Topic Maps paradigms
  - to make all of the advantages of the more familiar approaches available in combination with each other
  - to allow the knowledge output of existing systems and processes to be integrated without first having to modify them

All these requirements can be met with a hybrid approach that leverages the ability of internationally standard paradigms, notably the GROVE paradigm and the Topic Maps paradigm, to make all of the advantages of the more familiar approaches available in combination with each other, and to allow the knowledge output of existing systems and processes to be integrated without first having to modify them.



Components of a Hybrid Approach

- Existing approaches to refining data
  - Custom metadata
  - Full text retrieval
  - Web searching
  - Expert inferencing
- Topic Maps: to amalgamate knowledge
- GROVE paradigm: to add essentially unlimited scale and arbitrarily fine resolution of addressing, notation independence, a rigorous doctrine of information identity, and auditability

Custom metadata systems are popular and relatively foolproof ways of managing massive, valuable corpora of information as they are exploited in the contexts of relatively stable business models. The owner decides what kinds of information must be stored about each piece of information and/or subject, and creates a database that records this information. Retrieval is generally fast, and scale is limited only by current database technology. Custom metadata systems can also be quite sophisticated and complex.

Exploiting the data housed in a custom metadata system can be fast and easy, but data **preparation** in these systems can be a significant bottleneck. The people who prepare the data often must deeply understand the purpose and design of each kind of metadata, and they must understand the content and subjects that they are, in effect, filing. They must have specialized knowledge, skills, and abilities, and they must conscientiously work together in order to create a consistent product.

Precisely because internal consistency is at the heart of the ease with which the contents of custom metadata systems can be exploited, control over the nature of the metadata, and the world view on which the metadata schema is based, cannot be distributed.

Full-text retrieval

Full text retrieval systems are extremely powerful and useful for word-based searching. The time required to prepare the data for later retrieval is negligible. The scale on which full-text retrieval can be accomplished is very large. Since full text retrieval systems are not concerned with concepts, but only with the words used to express them, there is no problem of ontological diversity and/or incompatibility; there is virtually no ontology at all. The usefulness ofthe full - text retrieval paradigm does not depend on the idea that the metadata or data are consistent in any way.

Data preparation is not a bottleneck in full-text retrieval scenarios, but **data exploitation** is. Considerable time and effort may be required to extract the data that are relevant to a particular concept. Full text queries can result in uselessly large numbers of hits, not only because the queries cannot be made precisely, in terms of the subject of interest, but also becausemany "false positives" are regarded by the system as relevant only because of the vagaries of natural language, in which a single word or phrase may refer to many different concepts. Full text queries also often fail to return relevant material (i.e., they yield "false negatives"), again because of the vagaries of natural language, which often uses different words to refer to a single concept. Although it is often useful to do so, it is expensive and time-consuming for human beings to use full-text searching, and the comprehensiveness of the results cannot be assured.

Web search

Web-based searching is already a kind of hybrid. It incorporates full-text retrieval systems, and, unlike them, it offers the ability to distribute control over what is made available and how it is made findable. The Web can accommodate many warehouses, and it offers a unified address space within which certain kinds of information components -- including HTML and XML elements -- are uniformly addressable.

However, as currently deployed, the Web is unable to provide the needed transparency between the full range of evidence and the conclusions drawn fromit, basically because of two weaknesses in its addressing capabilities:

1. Only some components are provided with well-defined addressing services. The Web does not attempt to provide system-neutral addressing services for all components of all information, regardless of notation. Instead, the Web effectively privileges certain notations, such as XML itself, various XML vocabularies, and, of course, HTML. The Web's leadership have, at least up to the present time, not chosen to provide public mechanisms designed to enfranchise others with the ability to bring other notations under a consistent addressing umbrella which would provide the same basic addressing services for all notations that is currently provided for the XML notation bythe W3C DOM.

2. Current Web technology does not provide a way to determine whether two different addressing expressions (URIs) address the same informationcomponent. More generally, it lacks a rigorous concept of information identity.

In addition, Web technology is a suboptimal foundation for the support ofreliable collaborative maintenance of webs of n-directional hyperlinks, such as semantic networks. Substantial human effort is required to maintain web pages that contain links to independent web pages, and some significant fraction of this effort could be avoided if the Web could provide a platform capable of supporting auditable collaborative information management processes. (Of course, the Web is well suited as a **publishing** medium for semantic networks. The design requirements for knowledge **management** systems are necessarily different from the requirements for knowledge **publishing** systems.)

The Web is a work in progress. It seems inevitable that, at some future date, these weaknesses will be resolved. The Semantic Web initiative, for example, may ultimately result in an enhanced Web that can support reliable distributed authoring of reliable destributed semantic networks.

Expert inferencing

Expert systems technology can reasonably be expected to make certain aspects of the problem of integrating huge quantities of information tractable. The great strength of expert systems is their ability to maintain and leverage a consistent body of knowledge. When sufficient knowledge has been accumulated in an adequately-performing expert system, the system can be used to identify the subjects to which a given piece of new content may be relevant. Such a system can be used as a kind of knowledge reactor, perhaps as part of a knowledge refinery. Even in complex **data preparation** processes, it may replacehuman-hours with machine-minutes, creating "finding information" that, at least within its field of expertise and world view, is consistent. The resulting metadata can be exploited quickly and easily by both human and machine analysts,significantly amplifying their productivity.

Expert systems have weaknesses, too.

1. Building an expert system is time -consuming, exacting, highly skilled workthat only humans can do.

2. As in the case of human experts, there is no way to prove that a useful expert system is drawing accurate conclusions. The utility of such a system is precisely in its ability to give us answers that are otherwise unobtainable (or at least impractical to obtain).

3. Each expert system is limited to a single world view and ontology.

4. There are practical limits on the size of an expert system, i.e., the amount of expertise it can accumulate and still perform.

While it may seem obvious that all we need to do is to combine all ontologies and all of civilization's knowledge into a single expert system, there is little hope, at least for the immediate future, of accomplishing such a feat. Moreover, the value of attempting to combine all world views and ontologies is questionable, because the value of knowledge is directly related to an appreciation of its original context, which is always a specific world view. Incoming knowledge that may be of value for detecting terrorist threats would rarely emanate from a world view that is (somehow) the sum of all world views.

Therefore, expert inferencing systems are best used for analytical work done entirely within some ontology, and via an accumulated knowledge base that has been limited to a size that is consistent with a useful level of performance. In this role, they can refine data rapidly, by means of as much hardware as may be necessary to handle the daily avalanche of new information, so that rapid exploitation of the data becomes possible.

A hybrid approach

What's needed, then, is a way to make a useful integration of the work of many independent knowledge refineries, including both expert systems and humanbeings, simultaneously. The Topic Maps paradigm, with its ontology-independent merging model, is part of the answer. The GROVE paradigm adds essentially unlimited scale and arbitrarily fine resolution of addressing, notation independence, a rigorous concept of information identity, and auditability.

Adding the Topic Maps paradigm

The basic feature of the merging model of Topic Maps, and, indeed of the paradigm itself, is that a topic map -- the abstract thing for which an internationally standard syntax exists -- is a semantic network in which there is one node per subject, one subject per node, and all subjects can be explicit. There is no redundancy in a topic map.

Since there can be no more than one node for a given subject, everything that is known about that subject is represented as an assertion which is directlyconnected to that node. It doesn't matter how diverse are the ontologies within which the assertions are asserted (i.e., within which the things are known) about the subject. The ontological subjects are directly connected to the assertions to which they provide context; the assertions are self-describing.

The semantic network represented by a topic map is not necessarily internally consistent. Topic maps are merely amalgamations of knowledge of any size, that can include contributions from any number of sources, within any number of ontologies. As such, Topic Maps systems do not directly or inherently offer the inferencing power that is available in the context of a true expert system. However, and by the same token, Topic Maps systems are not subject to the same limitations as expert systems. They do not suffer from scaling limitations; their performance need not diminish, even when their size is large enough to integrate a significant fraction of the world's information. Adding knowledge to them is not a painstaking, time-consuming process in which all inconsistencies must be eliminated.

Nonetheless, a topic map is a semantic network in which an expert system or a human analyst can hope to detect patterns of terrorist behavior that would be impossible to detect without first integrating knowledge emanating from a large number of diverse sources. Topic maps can be used as a way to represent partially-refined, very large scale knowledge bases in which the outputs of many different refinement processes -- including refinement processes that depend on expert systems -- can be amalgamated. Further refinement processes that may also depend on expert systems can then be applied. These second-stage expert systems may embody knowledge of different ontologies, including ontologies that encompass combinations of other ontologies. In theory, further stages of refinement may also involve amalgamating information derived by combining the information refined in terms of these master ontologies. At some stage of refinement, combinations of facts that have been deemed to be threatening can be detected.

The general idea behind using topic maps to amalgamate knowledge is a simple and classic one: in order to solve an unmanageably large problem, subdivide it into many smaller, manageable ones, and construct a pyramid of progressive solutions that can yield an answer of manageable proportions. Neither an expert system nor an analyst can handle the scale and diversity of the overall problem space, but many expert systems and many analysts, operating at the same time, may get useful results if they can operate in concert, "on the same buss". The Topic Maps paradigm provides such a buss, and without compromising comprehensiveness or precision. It can be used to eliminate a class of bottlenecks, putting governments in a position to beat the terrorism forecasting problem into submission by "throwing hardware at it"; i.e. they can strike a balance between the quantity of dedicated hardware and the duration of elapsed integration time.

There is another benefit of the Topic Maps paradigm which may prove to be significant: the fact that it is an international standard which can be used to competitive advantage in many kinds of knowledge-oriented businesses, including legal and technical publishing, financial services, healthcare, marketing, etc. To the extent that businesses use the paradigm as a knowledge management tool, and to the extent that businesses make their relevant topic maps available to governments for terrorism forecasting purposes, knowledge refinement expense canbe avoided.

Adding the GROVE paradigm

In order to be used at large scales, with diverse information notations, with full auditability of the knowledge refinement processes, the Topic Maps paradigm needs an underlying addressing paradigm that is equal to the challenge. The success of the Web demonstrates the huge scale that can be achieved using the client-server model in combination with a single address space. However, the Web was not designed to meet the requirements of a large-scale terrorism forecasting system, and it lacks the following required abilities:

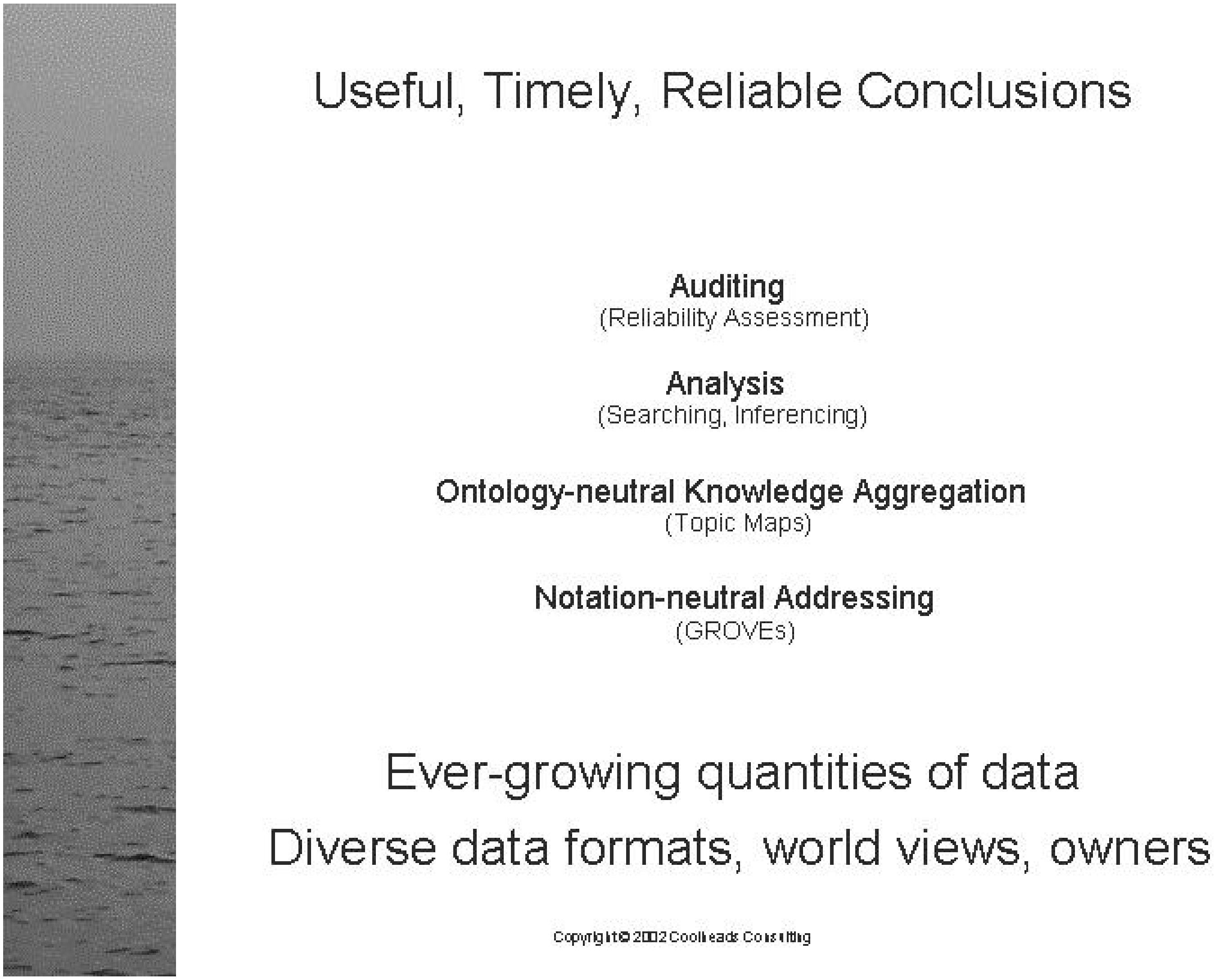
1. The ability to effectively link every component of information to all of the things that reference it. (Sometimes this kind of feature is called "back linking".) This feature is key to providing fully auditable information handling, and the GROVE paradigm is designed to support it.

2. The ability to address all the contents of all kinds of repositories, databases, and knowledge bases, and all the components of their contents, despite the diversity of their notations, and despite the fact that some of these repositories and notations haven't been invented yet, within a single address space. While it's true that the Web's URI mechanism allows all kinds of information to be embedded within addressing expressions, including information that can be interpreted by repositories, databases, and knowledgebases as addressing expressions, this by itself does not meet the requirement. Addressing expressions that contain instructions to software are not protectedfrom technological change. The value of a significant investment can (and often does) depend on an addressing expression. If an addressing expression contains instructions that can only be understood by a certain piece ofaddress-resolution software, then the investment represented by that addressing expression will be lost when that software is changed. Knowledge is lost. Auditability cannot be reliably supported over the long term. By using the GROVE paradigm, addressing expressions can be protected from technology change, because GROVE-based addressing expressions invoke only the intrinsic properties of the content being addressed. Even if the software (the "GROVE builder") that exposes those properties becomes unavailable, its addressing specifications remain the subject of a formal, machine-readable document (a "property set" conforming to an internationally standardized DTD), and the software can be replaced with a reliable substitute.

3. With GROVES, addressing expressions remain just as powerful and flexible asURIs, but they are not as opaque as URIs. Instead, they are automatically interpretable and manipulable, to a significant extent, regardless of differences between the various addressable data structures and interchange notations, and regardless of the manipulator's understanding of the semantics being conveyed by the data and its structure.

4. Even more fundamentally, the ability to integrate knowledge emanating from distributed sources depends on the ability to use pieces of addressable content as binding points for subjects. These pieces of addressable content must have identity. It must be possible to determine whether two different addressing expressions address the same piece of content.

Web technology currently does not support such determinations of identity. The GROVE paradigm provides each component of each information resource with a single strong and unambiguous identity. Given two addressing expressions for information under their control, GROVE-based servers can report whether they address the same piece of information.



Useful, Timely, Reliable Conclusions

Auditing
(Reliability Assessment)

Analysis
(Searching, Inferencing)

Ontology-neutral Knowledge Aggregation
(Topic Maps)

Notation-neutral Addressing
(GROVEs)

Ever-growing quantities of data

Diverse data formats, world views, owners

This paper summarizes the results of a series of discussions held at Coolheads Consulting. The contributors included Michel Biezunski, Peter Newcomb and Victoria T. Newcomb.